Machine Learning and Statistics in Genetics and Genomics

I: Course Overview and Introduction to Probability Theory

Christoph Lippert

Microsoft Research
eScience group

Los Angeles , USA

Current topics in computational biology
UCLA
Winter quarter 2014

Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications are not limited to statistical genetics: Machine Learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
  - Overview on probabilistic modeling
  - Key concepts
  - Focus on Applications in statistical genetics

Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications are not limited to statistical genetics: Machine Learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
  - Overview on probabilistic modeling
  - Key concepts
  - Focus on Applications in statistical genetics

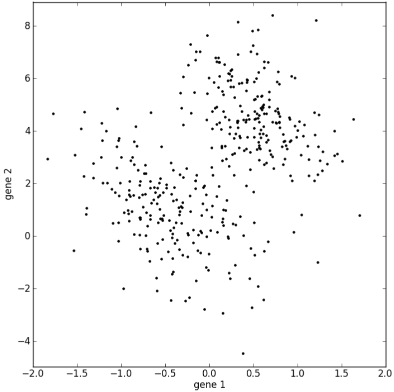
Why probabilistic modeling?

- Inferences from data are intrinsically uncertain.
- Probability theory: model uncertainty instead of ignoring it!
- Applications are not limited to statistical genetics: Machine Learning, Data Mining, Pattern Recognition, etc.
- Goal of this part of the course
  - Overview on probabilistic modeling
  - Key concepts
  - Focus on Applications in statistical genetics

Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. **Is** gene 1 co-expressed with gene 2?
  - Probabilistic models → probability theory
  - This course: linear models (and kernel methods)
- Is this dependence significant?
  - Statistical testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



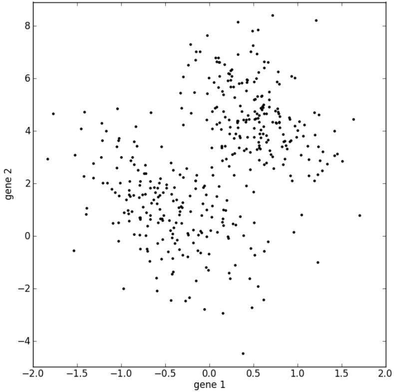
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models → probability theory
  - This course: linear models (and kernel methods)

$$\text{gene}_2 = c + \text{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - Statistical testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



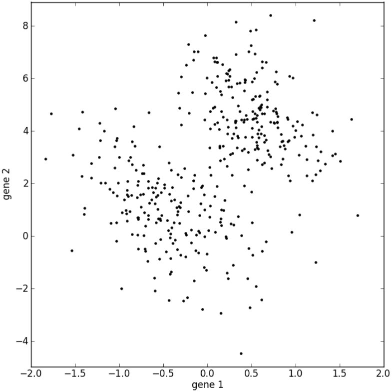
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\text{gene}_2 = c + \text{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - Statistical testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



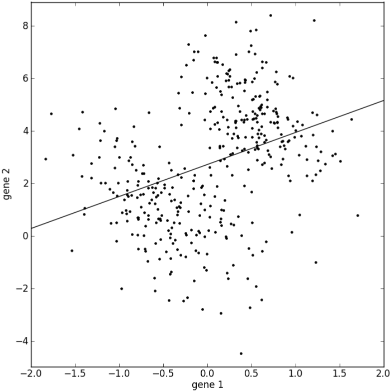
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\text{gene}_2 = c + \text{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



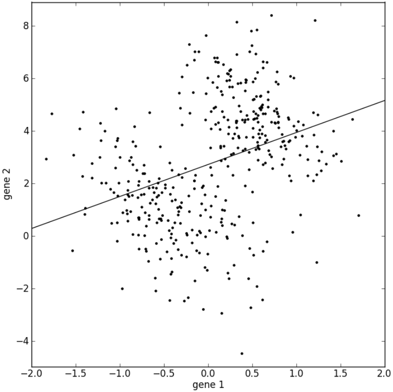
Why probabilistic modeling? Example

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\mathrm{gene}_2 = c + \mathrm{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?



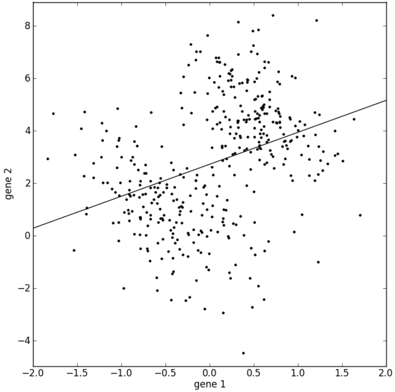
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\mathrm{gene}_2 = c + \mathrm{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders

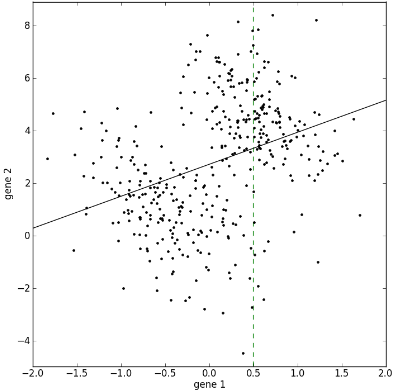


Why probabilistic modeling? Example

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\mathrm{gene}_2 = c + \mathrm{gene}_1 \cdot \color{red}{\beta} + \epsilon$$



- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?

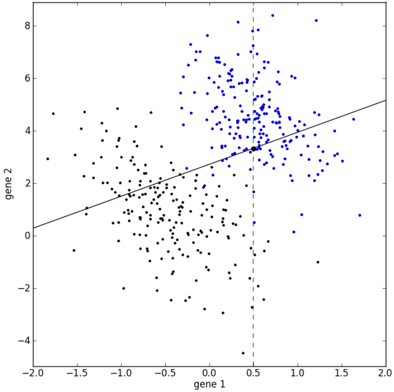
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\text{gene}_2 = c + \text{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



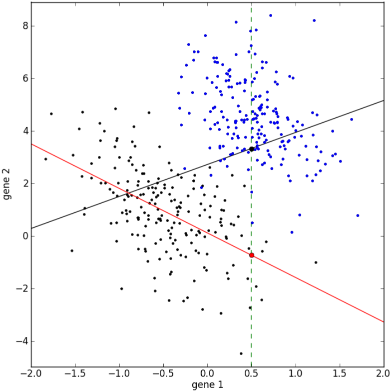
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\text{gene}_2 = c + \text{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



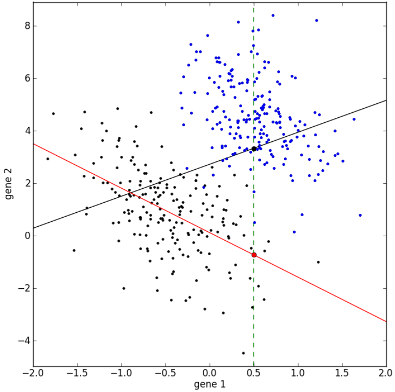
Why probabilistic modeling? Example

- Genes measured in yeast
- e.g. Is gene 1 co-expressed with gene 2?
  - Probabilistic models $\rightarrow$ probability theory
  - This course: linear models (and kernel methods)

$$\text{gene}_2 = c + \text{gene}_1 \cdot \beta + \epsilon$$

- Is this dependence significant?
  - **Statistical** testing
- Can I **predict** the level of gene2 observing gene1?

- Take **known** covariates into account
- **Estimate hidden** covariates/confounders



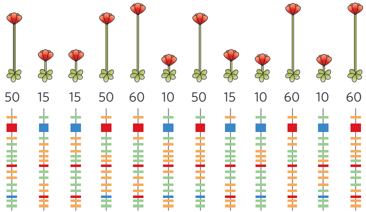
Why probabilistic modeling?

Example: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- Phenotypes for the same individuals
  - e.g.: disease, height, gene-expression, ...

- Try to find genetic markers, that explain the variance in the phenotype.

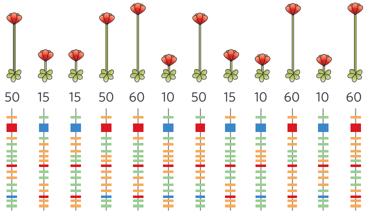


Why probabilistic modeling?

Example: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- Phenotypes for the same individuals
  - e.g.: disease, height, gene-expression, ...



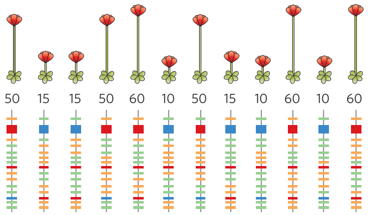
- Try to find genetic markers, that explain the variance in the phenotype.

Why probabilistic modeling?

Example: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- Phenotypes for the same individuals
  - e.g.: disease, height, gene-expression, ...



- Try to find genetic markers, that explain the variance in the phenotype.

Why probabilistic modeling?

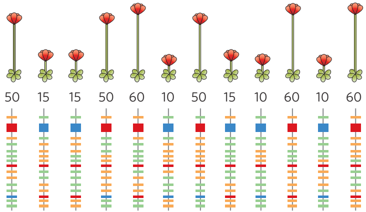
Example: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- Phenotypes for the same individuals
  - e.g.: disease, height, gene-expression, ...

Goal:

- Try to find genetic markers, that explain the variance in the phenotype.



Why probabilistic modeling?

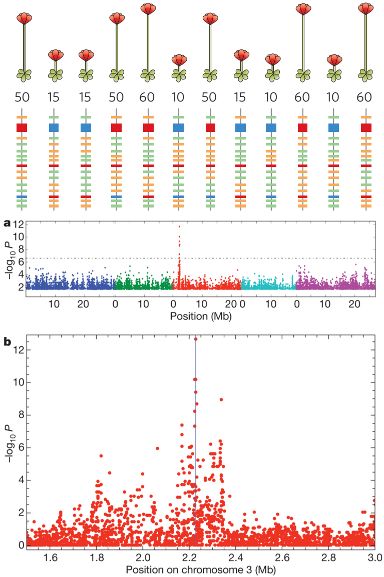
Example: Genome-wide association studies

Given:

- Genetics for multiple individuals
  - e.g.: Single nucleotide polymorphisms (SNPs), microsatelite markers, ...
- Phenotypes for the same individuals
  - e.g.: disease, height, gene-expression, ...

Goal:

- Try to find genetic markers, that explain the variance in the phenotype.

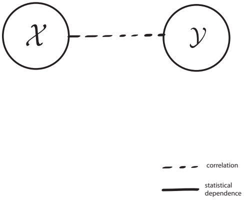


Why probabilistic modeling?

Example: Genome-wide association studies - continued

*In statistics, association is any relationship between two measured quantities that renders them statistically dependent.**

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure



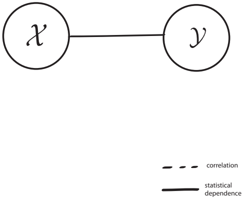
*Oxford Dictionary of Statistics

Why probabilistic modeling?

Example: Genome-wide association studies - continued

*In statistics, association is any relationship between two measured quantities that renders them statistically dependent.**

- Direct association
- Indirect association



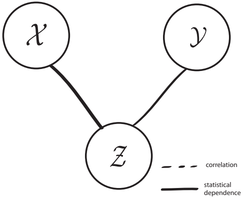
*Oxford Dictionary of Statistics

Why probabilistic modeling?

Example: Genome-wide association studies - continued

*In statistics, association is any relationship between two measured quantities that renders them statistically dependent.**

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure



*Oxford Dictionary of Statistics

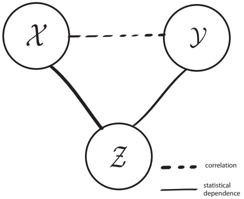
Why probabilistic modeling?

Example: Genome-wide association studies - continued

*In statistics, association is any relationship between two measured quantities that renders them statistically dependent.**

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure

*Oxford Dictionary of Statistics

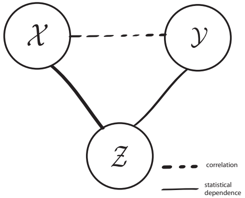


Why probabilistic modeling?

Example: Genome-wide association studies - continued

*In statistics, association is any relationship between two measured quantities that renders them statistically dependent.**

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure



*Oxford Dictionary of Statistics

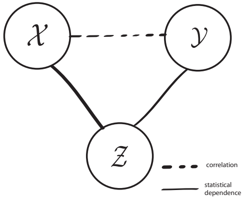
Why probabilistic modeling?

Example: Genome-wide association studies - continued

*In statistics, association is any relationship between two measured quantities that renders them statistically dependent.**

- Direct association
- Indirect association
  - Can be beneficial
    e.g.: Linkage
  - Can be harmful
    e.g.: Population structure

*Oxford Dictionary of Statistics



Further reading, useful material

- Christopher M. Bishop: Pattern Recognition and Machine learning.
  - Good background, covers most of the machine learning used in this course and much more!
  - Substantial parts of this tutorial borrow figures and ideas from this book.
- David J.C. MacKay: Information Theory, Learning and Inference
  - Very worthwhile reading, not quite the same quality of overlap with the lecture synopsis.
  - Freely available online.

Course structure

- Probability Theory
  - Rules of probability calculus
  - Distributions
- Linear models (statistics)
  - Linear regression
  - Parameter estimations
  - Statistical testing
  - Regularization (ridge, Lasso)
  - Random effects models
  - Linear mixed models
- Latent variable models
  - Principle components analysis (PCA)
  - Mixture models
- Kernel methods
  - Introduction to kernels
  - Non-parametric regression (Gaussian Process)
  - Non-linear PCA models (kernel PCA, GPLVM)
  - Multivariate regression

Course structure

- Probability Theory
  - Rules of probability calculus
  - Distributions
- Linear models (statistics)
  - Linear regression
  - Parameter estimations
  - Statistical testing
  - Regularization (ridge, Lasso)
  - Random effects models
  - Linear mixed models
- Latent variable models
  - Principle components analysis (PCA)
  - Mixture models
- Kernel methods
  - Introduction to kernels
  - Non-parametric regression (Gaussian Process)
  - Non-linear PCA models (kernel PCA, GPLVM)
  - Multivariate regression

Course structure

- Probability Theory
  - Rules of probability calculus
  - Distributions
- Linear models (statistics)
  - Linear regression
  - Parameter estimations
  - Statistical testing
  - Regularization (ridge, Lasso)
  - Random effects models
  - Linear mixed models

- Latent variable models
  - Principle components analysis (PCA)
  - Mixture models
- Kernel methods
  - Introduction to kernels
  - Non-parametric regression (Gaussian Process)
  - Non-linear PCA models (kernel PCA, GPLVM)
  - Multivariate regression

Course structure

- Probability Theory
  - Rules of probability calculus
  - Distributions
- Linear models (statistics)
  - Linear regression
  - Parameter estimations
  - Statistical testing
  - Regularization (ridge, Lasso)
  - Random effects models
  - Linear mixed models
- Latent variable models
  - Principle components analysis (PCA)
  - Mixture models
- Kernel methods
  - Introduction to kernels
  - Non-parametric regression (Gaussian Process)
  - Non-linear PCA models (kernel PCA, GPLVM)
  - Multivariate regression

Course structure

- Probability Theory
  - Rules of probability calculus
  - Distributions
- Linear models (statistics)
  - Linear regression
  - Parameter estimations
  - Statistical testing
  - Regularization (ridge, Lasso)
  - Random effects models
  - Linear mixed models
- Latent variable models
  - Principle components analysis (PCA)
  - Mixture models
- Kernel methods
  - Introduction to kernels
  - Non-parametric regression (Gaussian Process)
  - Non-linear PCA models (kernel PCA, GPLVM)
  - Multivariate regression

Outline

Outline

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment

- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

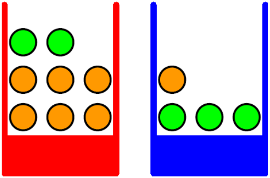
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

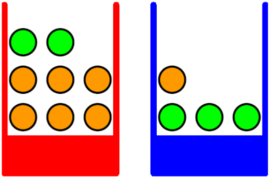
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment

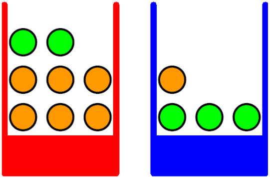


experiment	pick a box and then take a ball at random	coin flip	gene expression measurement
sample space Ω, $P(\Omega) = 1$	$\Omega = \{RG, RO, BG, BO\}$	$\Omega = \{H, T\}$	$\Omega =]-\infty, \infty[$
event Subsets of Ω	$A = \{RG, RO\}$, $B = \{RO, BO\}$	$A = \{H\}$, $B = \{T\}$	$A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

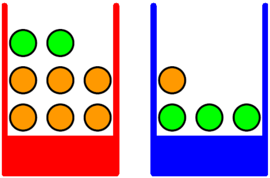
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

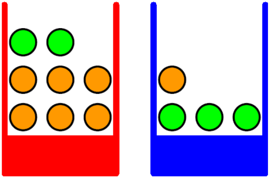
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

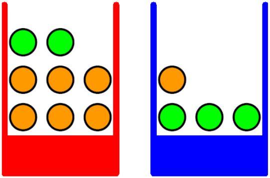
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

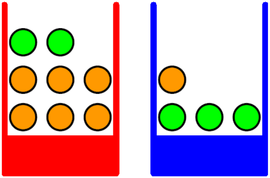
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

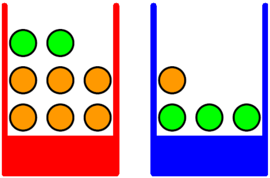
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

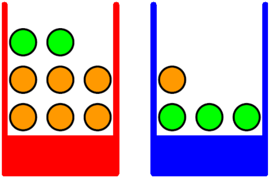
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

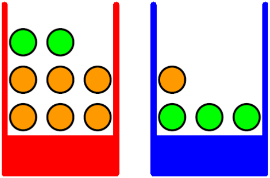
- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Probabilities

- Probabilities describe likeliness of the outcomes of an experiment



- experiment
- sample space Ω, $P(\Omega) = 1$
- event
  Subsets of Ω

- pick a box and then take a ball at random
- $\Omega = \{RG, RO, BG, BO\}$
- $A = \{RG, RO\}$, $B = \{RO, BO\}$

- coin flip
- $\Omega = \{H, T\}$
- $A = \{H\}$, $B = \{T\}$

- gene expression measurement
- $\Omega =]-\infty, \infty[$
- $A =]-\infty, 3]$, $B = \{2\}$

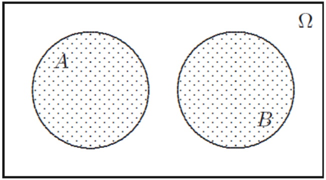
(C.M. Bishop, Pattern Recognition and Machine Learning)

Probability function

- Probability functions are non-negative, $P(A) \geq 0$
- If A and B are *disjoint*
  - $P(A \cup B) = P(A) + P(B)$ (union)
  - $P(A \cap B) = 0$ (intersection)
- Probabilities sum to 1 over union of all possible *disjoint* events $A_1 \cup A_2 \cup \ldots$
- $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots = P(\Omega) = 1$

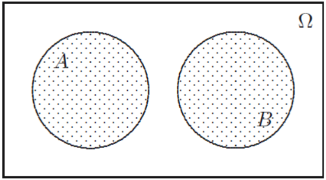
Probability function

- Probability functions are non-negative, $P(A) \geq 0$
- If A and B are *disjoint*
  - $P(A \cup B) = P(A) + P(B)$ (union)
  - $P(A \cap B) = 0$ (intersection)
- Probabilities sum to 1 over union of all possible *disjoint* events $A_1 \cup A_2 \cup \ldots$
- $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots = P(\Omega) = 1$



Probability function

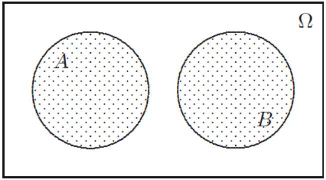
- Probability functions are non-negative, $P(A) \geq 0$
- If A and B are *disjoint*
  - $P(A \cup B) = P(A) + P(B)$ (union)
  - $P(A \cap B) = 0$ (intersection)
- Probabilities sum to 1 over union of all possible *disjoint* events $A_1 \cup A_2 \cup \ldots$



- $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots = P(\Omega) = 1$

Probability function

- Probability functions are non-negative, $P(A) \geq 0$
- If A and B are *disjoint*
  - $P(A \cup B) = P(A) + P(B)$ (union)
  - $P(A \cap B) = 0$ (intersection)
- Probabilities sum to 1 over union of all possible *disjoint* events $A_1 \cup A_2 \cup \ldots$

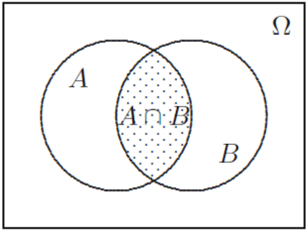


- $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \cdots = P(\Omega) = 1$

Intersection and Union

- Intersection $A \cap B$

$$\begin{aligned} & P(A \cap B) \\ = & P(A) + P(B) - P(A \cup B) \end{aligned}$$



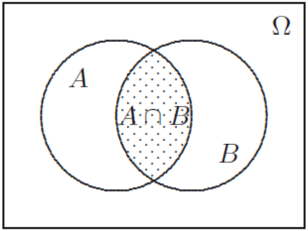
- Union $A \cup B$

$$\begin{aligned} & P(A \cup B) \\ = & P(A) + P(B) - P(A \cap B) \end{aligned}$$

Intersection and Union

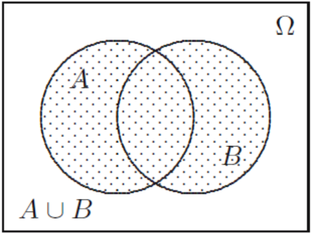
- Intersection $A \cap B$

$$
\begin{aligned}
&P(A \cap B) \\
=&P(A) + P(B) - P(A \cup B)
\end{aligned}
$$



- Union $A \cup B$

$$
\begin{aligned}
&P(A \cup B) \\
=&P(A) + P(B) - P(A \cap B)
\end{aligned}
$$

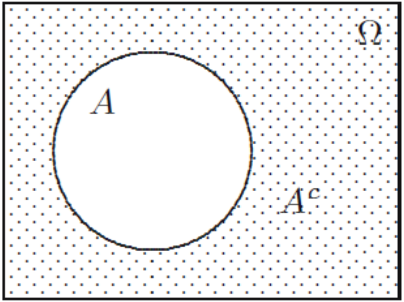


Complement and DeMorgan's Laws

- The complement of A is denoted by A^c
- $\Omega^c = \emptyset$

$$P(A^c) = 1 - P(A)$$
$$P(A \cap A^c) = 0$$
$$P(A \cup A^c) = 1$$

- DeMorgan's Laws:
- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

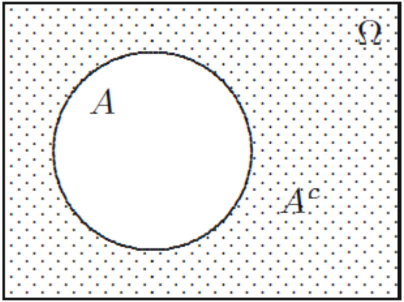


Complement and DeMorgan's Laws

- The complement of A is denoted by A^c
- $\Omega^c = \emptyset$

$$P(A^c) = 1 - P(A)$$
$$P(A \cap A^c) = 0$$
$$P(A \cup A^c) = 1$$

- DeMorgan's Laws:
- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$

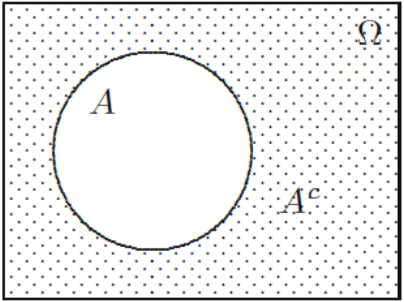


Complement and DeMorgan's Laws

- The complement of A is denoted by A^c
- $\Omega^c = \emptyset$

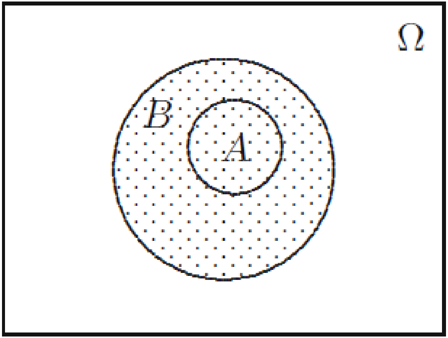
$$P(A^c) = 1 - P(A)$$
$$P(A \cap A^c) = 0$$
$$P(A \cup A^c) = 1$$

- DeMorgan's Laws:
- $(A \cup B)^c = A^c \cap B^c$
- $(A \cap B)^c = A^c \cup B^c$



Events

- If A is a subset of B
- $A \cup B = B$
- $P(A \cup B) = P(B)$
- $A \cap B = A$
- $P(A \cap B) = P(A)$



Products of sample spaces

- Typically we don't perform only a single experiment
- Repeated experiments
  - Flip a coin N times
  - Measure a phenotype at different time points
- Multiple experiments
  - Measure expression of multiple genes
- Sample space is product of sample spaces
  - $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$
  - number of elements multiply

- Experiments can be independent
  - Flip a coin twice
  - $P(H, H) = P(H)^2$
- or dependent
  - Dependence of measurements over time
  - Two genes that are co-regulated
  - $P(g_1 = x, g_2 = y) \neq P(g_1 = x) \cdot P(g_2 = y)$

Products of sample spaces

- Typically we don't perform only a single experiment
- Repeated experiments
  - Flip a coin N times
  - Measure a phenotype at different time points
- Multiple experiments
  - Measure expression of multiple genes
- Sample space is product of sample spaces
  - $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$
  - number of elements multiply

- Experiments can be independent
  - Flip a coin twice
  - $P(H, H) = P(H)^2$
- or dependent
  - Dependence of measurements over time
  - Two genes that are co-regulated
  - $P(g_1 = x, g_2 = y) \neq P(g_1 = x) \cdot P(g_2 = y)$

Products of sample spaces

- Typically we don't perform only a single experiment
- Repeated experiments
  - Flip a coin N times
  - Measure a phenotype at different time points
- Multiple experiments
  - Measure expression of multiple genes
- Sample space is product of sample spaces
  - $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$
  - number of elements multiply
- Experiments can be independent
  - Flip a coin twice
  - $P(H, H) = P(H)^2$
- or dependent
  - Dependence of measurements over time
  - Two genes that are co-regulated
  - $P(g_1 = x, g_2 = y) \neq P(g_1 = x) \cdot P(g_2 = y)$

Products of sample spaces

- Typically we don't perform only a single experiment
- Repeated experiments
  - Flip a coin N times
  - Measure a phenotype at different time points
- Multiple experiments
  - Measure expression of multiple genes
- Sample space is product of sample spaces
  - $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$
  - number of elements multiply

- Experiments can be independent
  - Flip a coin twice
  - $P(H, H) = P(H)^2$
- or dependent
  - Dependence of measurements over time
  - Two genes that are co-regulated
  - $P(g_1 = x, g_2 = y) \neq P(g_1 = x) \cdot P(g_2 = y)$

Products of sample spaces

- Typically we don't perform only a single experiment
- Repeated experiments
  - Flip a coin N times
  - Measure a phenotype at different time points
- Multiple experiments
  - Measure expression of multiple genes
- Sample space is product of sample spaces
  - $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$
  - number of elements multiply

- Experiments can be independent
  - Flip a coin twice
  - $P(H, H) = P(H)^2$
- or dependent
  - Dependence of measurements over time
  - Two genes that are co-regulated
  - $P(g_1 = x, g_2 = y) \neq P(g_1 = x) \cdot P(g_2 = y)$

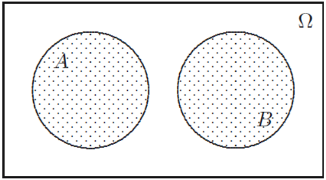
Products of sample spaces

- Typically we don't perform only a single experiment
- Repeated experiments
  - Flip a coin N times
  - Measure a phenotype at different time points
- Multiple experiments
  - Measure expression of multiple genes
- Sample space is product of sample spaces
  - $\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_N$
  - number of elements multiply

- Experiments can be independent
  - Flip a coin twice
  - $P(H, H) = P(H)^2$
- or dependent
  - Dependence of measurements over time
  - Two genes that are co-regulated
  - $P(g_1 = x, g_2 = y) \neq P(g_1 = x) \cdot P(g_2 = y)$

Conditional probability

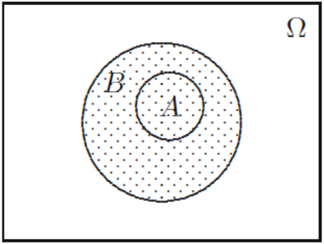
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

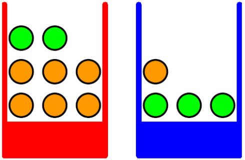
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements

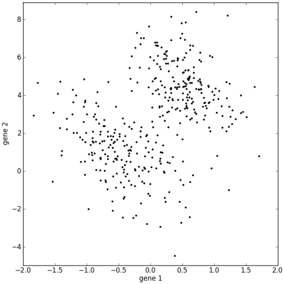


(C.M. Bishop, Pattern Recognition and Machine Learning)

- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

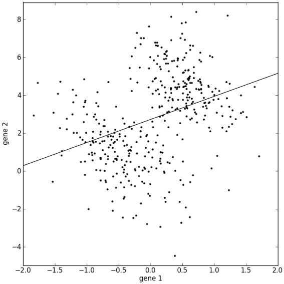
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

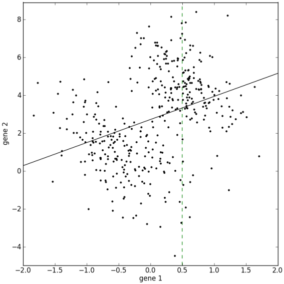
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

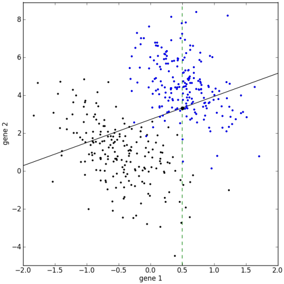
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

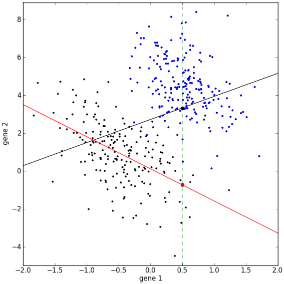
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

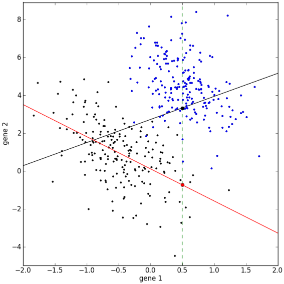
- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements



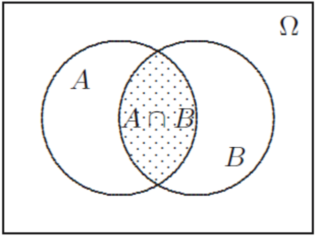
- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

Conditional probability

- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements

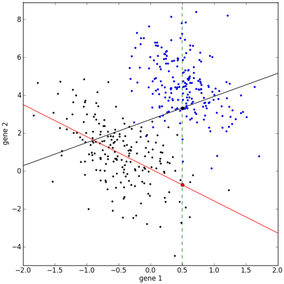


- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)

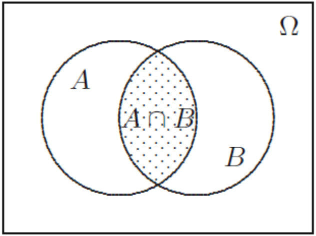


Conditional probability

- Some times occurrence of one event yields information about another one
- disjoint events
- subsets
- dependent measurements

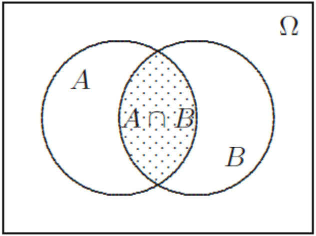


- $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)}$
- $P(A \cap B) = P(A \mid B) \cdot P(B)$ (Product rule)



Independence

- The three following statements are equivalent and imply independence of A and B:
  - $P(A \mid B) = P(A)$,
  - $P(B \mid A) = P(B)$,
  - $P(A \cap B) = P(A) \cdot P(B)$.



Random variables

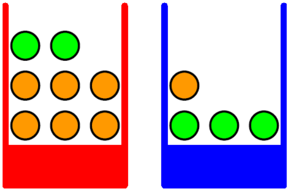
- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

- Number of orange balls in N trials
(discrete)
$\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T
(discrete)
$\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete)
$\mathcal{X} = \{2, 3, \ldots, 12\}$
- Gene expression at time t (continuous)
$\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous)
$\mathcal{X} = \mathbb{R}$

Random variables

- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

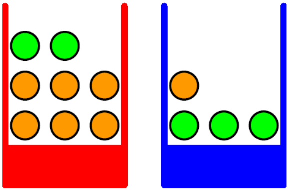
- Number of orange balls in N trials (discrete) $\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T (discrete) $\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete) $\mathcal{X} = \{2, 3, \ldots, 12\}$
- Gene expression at time t (continuous) $\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous) $\mathcal{X} = \mathbb{R}$



Random variables

- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

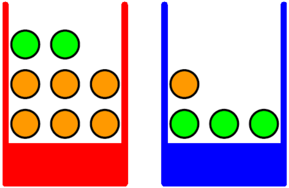
- Number of orange balls in N trials (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete)
  $\mathcal{X} = \{2, 3, \ldots, 12\}$
- Gene expression at time t (continuous)
  $\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous)
  $\mathcal{X} = \mathbb{R}$



Random variables

- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

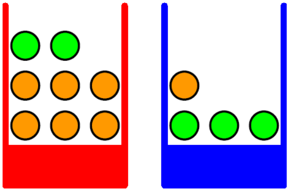
- Number of orange balls in N trials (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete)
  $\mathcal{X} = \{2, 3, \dots, 12\}$
- Gene expression at time t (continuous)
  $\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous)
  $\mathcal{X} = \mathbb{R}$



Random variables

- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

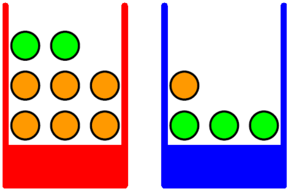
- Number of orange balls in N trials (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete)
  $\mathcal{X} = \{2, 3, \ldots, 12\}$
- Gene expression at time t (continuous)
  $\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous)
  $\mathcal{X} = \mathbb{R}$



Random variables

- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

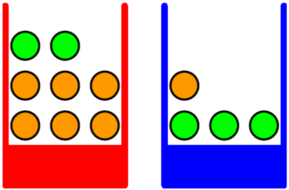
- Number of orange balls in N trials (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T (discrete)
  $\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete)
  $\mathcal{X} = \{2, 3, \ldots, 12\}$
- Gene expression at time t (continuous)
  $\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous)
  $\mathcal{X} = \mathbb{R}$



Random variables

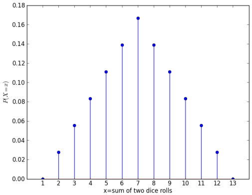
- Alternatively to defining sets of events we can define random variables of interest.
- A random variables X is defined over a set of possible values $\mathcal{X}$

- Number of orange balls in N trials (discrete) $\mathcal{X} = \mathbb{N}_0^+$
- Number of H coin flips before first T (discrete) $\mathcal{X} = \mathbb{N}_0^+$
- Sum of two dice rolls (discrete) $\mathcal{X} = \{2, 3, \ldots, 12\}$
- Gene expression at time t (continuous) $\mathcal{X} = \mathbb{R}$
- Average gene-expression measurement over N samples (continuous) $\mathcal{X} = \mathbb{R}$

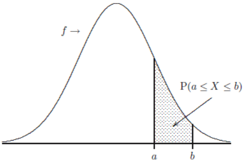


Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



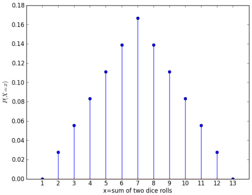
$$\sum_{x \in \mathcal{X}} p(x) = 1$$



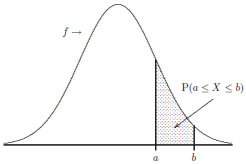
$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



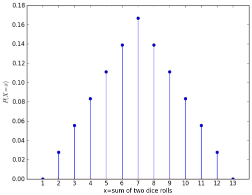
$$\sum_{x \in \mathcal{X}} p(x) = 1$$



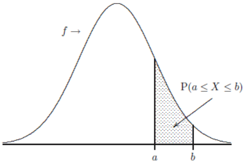
$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



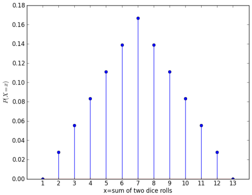
$$\sum_{x \in \mathcal{X}} p(x) = 1$$



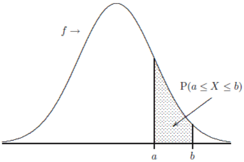
$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



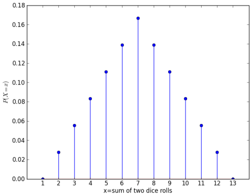
$$\sum_{x \in \mathcal{X}} p(x) = 1$$



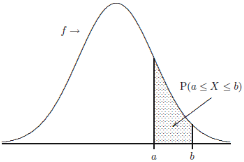
$$\int_{x \in \mathcal{X}} p(x)\mathrm{d}x = 1$$

Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



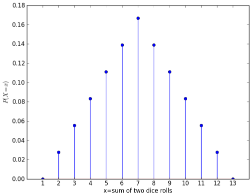
$$\sum_{x \in \mathcal{X}} p(x) = 1$$



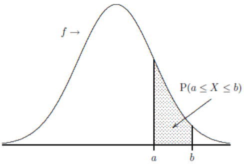
$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



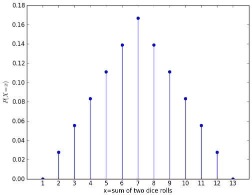
$$\sum_{x \in \mathcal{X}} p(x) = 1$$



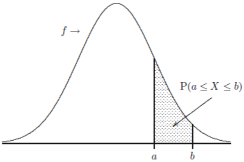
$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

Probabilities and random variables

- Let X be a random variable, defined over a set $\mathcal{X}$ or measurable space.
- $P(X = x)$ denotes the probability that X takes value x, short $p(x)$.
  - Probability mass function (discrete)
  - Probability density function (continuous)
  - Probabilities are non-negative, $P(X = x) \geq 0$
  - Probabilities sum to one



$$\sum_{x \in \mathcal{X}} p(x) = 1$$



$$\int_{x \in \mathcal{X}} p(x) \mathrm{d}x = 1$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

- Measures average squared deviation fro
the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N} \left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] = \sum_{x\in\mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}} \left(x - \mathbb{E}\left[X\right]\right)^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}[X]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}[X] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

- Measures average squared deviation fro the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{x\in\mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}} (x - \mathbb{E}[X])^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}[X]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}[X] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

- Measures average squared deviation fro the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{x\in\mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}} (x - \mathbb{E}[X])^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

- Measures average squared deviation from the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N}\left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] = \sum_{x\in\mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}} \left(x - \mathbb{E}\left[X\right]\right)^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x \in \mathcal{X}} x \cdot p(x) \mathrm{d}x$$

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N} \sum_{n=1}^{N} (x_n - \bar{X})^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \sum_{x \in \mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x \in \mathcal{X}} (x - \mathbb{E}\left[X\right])^2 \cdot p(x) \mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fro
the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{X})^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \sum_{x\in\mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}} (x - \mathbb{E}\left[X\right])^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{X})^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] = \sum_{x\in\mathcal{X}}(x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}}(x - \mathbb{E}\left[X\right])^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x \in \mathcal{X}} x \cdot p(x) \mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N} \left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] = \sum_{x \in \mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x \in \mathcal{X}} \left(x - \mathbb{E}\left[X\right]\right)^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}[X]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}[X] = \int_{x \in \mathcal{X}} x \cdot p(x) \mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N} \sum_{n=1}^{N} \left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{x \in \mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 \cdot p(x) \mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}\left[X\right]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}\left[X\right] = \sum_{x \in \mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}\left[X\right] = \int_{x \in \mathcal{X}} x \cdot p(x) \mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N} \sum_{n=1}^{N} \left(x_n - \bar{X}\right)^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right] = \sum_{x \in \mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x \in \mathcal{X}} \left(x - \mathbb{E}\left[X\right]\right)^2 \cdot p(x) \mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}[X]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}[X] = \sum_{x\in\mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}[X] = \int_{x\in\mathcal{X}} x \cdot p(x)\mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N}\sum_{n=1}^{N} (x_n - \bar{X})^2$$

- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{x\in\mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x\in\mathcal{X}} (x - \mathbb{E}[X])^2 \cdot p(x)\mathrm{d}x$$

Expected values and variances

Moments

Expected value

- Average value of the random variable X
- sample mean $\bar{X}$ of a data sample drawn from $p(x)$.

$$\bar{X} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

- Expected value $\mathbb{E}[X]$ is the first moment of $P(X)$
- discrete

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot p(x)$$

- continuous

$$\mathbb{E}[X] = \int_{x \in \mathcal{X}} x \cdot p(x) \mathrm{d}x$$

Variance σ^2

- Measures average squared deviation fror the mean of X.
- sample variance of a data sample drawn from $p(x)$.

$$\frac{1}{N} \sum_{n=1}^{N} \left(x_n - \bar{X}\right)^2$$

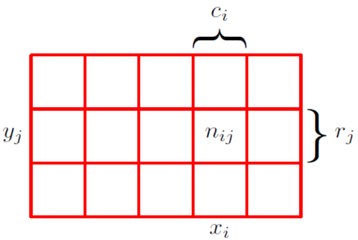
- Second centralized moment of X
- square of the *standard deviation* σ
- discrete

$$\sigma^2(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \sum_{x \in \mathcal{X}} (x -$$

- continuous

$$\sigma^2(X) = \int_{x \in \mathcal{X}} (x - \mathbb{E}[X])^2 \cdot p(x) \mathrm{d}x$$

Distributions of multiple random variables



Joint Probability

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Marginal Probability

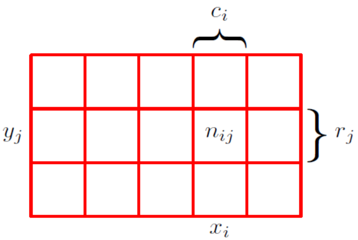
$$P(X = x_i) = \frac{c_i}{N}$$

Conditional Probability

$$P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Distributions of multiple random variables



Marginal Probability

$$P(X = x_i) = \frac{c_i}{N}$$

Conditional Probability

$$P(Y = y_j \mid X = x_i) = \frac{n_{ij}}{c_i}$$

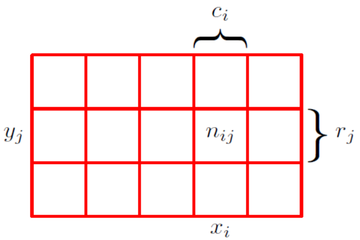
Product Rule

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= P(Y = y_j \mid X = x_i) P(X = x_i)$$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Distributions of multiple random variables



Sum Rule

$$P(X = x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{L} n_{ij}$$

$$= \sum_j P(X = x_i, Y = y_j)$$

Product Rule

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N}$$

$$= P(Y = y_j \mid X = x_i)P(X = x_i)$$

(C.M. Bishop, Pattern Recognition and Machine Learning)

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

- Entropy is a vector-valued function (input is a probability distribution)

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

- Entropy is a vector-valued function (input is a probability distribution)

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

- Entropy is a vector-valued function (input is a probability distribution)

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

- Entropy is a vector-valued function (input is a probability distribution)

Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

- Entropy is a vector-valued function (input is a probability distribution)

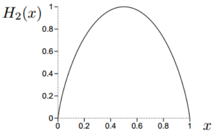
Information and Entropy

- Information is the reduction of uncertainty.
- Entropy $H(X)$ is the quantitative description of uncertainty
  - $H(X) = 0$: certainty about X.
  - $H(X)$ maximal if all possibilities are equal probable.
  - Uncertainty and information are additive.
- These conditions are fulfilled by the entropy function:

$$H(X) = -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x)$$

- Entropy is a vector-valued function (input is a probability distribution)

example:
binary entropy function



(D. MacKay, Information Theory, Inference, and Learning Algorithms)

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{\left(-\log P(X = x)\right)}_{\text{surprise}}$$

- Conditional entropy of X given $Y = y$

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y)$$

- Conditional entropy of X given Y is the *average* (over Y) conditional entropy of X given $Y = y$

$$H(X \mid Y) = \sum_{y \in \mathcal{Y}} P(Y = y) \left(-\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y) \right)$$

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X=x) \underbrace{\left(-\log P(X=x)\right)}_{\text{surprise}}$$

- Conditional entropy of X given $Y = y$

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y)$$

- Conditional entropy of X given Y is the *average* (over Y) conditional entropy of X given $Y = y$

$$H(X \mid Y) = \sum_{y \in \mathcal{Y}} P(Y = y) \left(-\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y) \right)$$

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{(-\log P(X = x))}_{\text{surprise}}$$

- Conditional entropy of X given $Y = y$

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y)$$

- Conditional entropy of X given Y is the *average* (over Y) conditional entropy of X given $Y = y$

$$H(X \mid Y) = \sum_{y \in \mathcal{Y}} P(Y = y) \left(-\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y) \right)$$

Definitions related to entropy and information

- Entropy is the average surprise

$$H(X) = \sum_{x \in \mathcal{X}} P(X = x) \underbrace{(-\log P(X = x))}_{\text{surprise}}$$

- Conditional entropy of X given $Y = y$

$$H(X \mid Y = y) = -\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y)$$

- Conditional entropy of X given Y is the *average* (over Y) conditional entropy of X given $Y = y$

$$\begin{aligned} H(X \mid Y) &= \sum_{y \in \mathcal{Y}} P(Y = y) \left(-\sum_{x \in \mathcal{X}} P(X = x \mid Y = y) \log P(X = x \mid Y = y) \right) \\ &= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x, Y = y) \log P(X = x \mid Y = y) \end{aligned}$$

Definitions related to entropy and information

- Chain rule

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

- Mutual information

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

Definitions related to entropy and information

- Chain rule

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

- Mutual information

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

  - $I(X;Y) = I(Y;X)$
  - average reduction in uncertainty about X when learning value of Y (and vice versa)

Definitions related to entropy and information

- Chain rule

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

- Mutual information

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

  - $I(X;Y) = I(Y;X)$
  - average reduction in uncertainty about X when learning value of Y (and vice versa)

Definitions related to entropy and information

- Chain rule

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

- Mutual information

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

  - $I(X;Y) = I(Y;X)$
  - average reduction in uncertainty about X when learning value of Y (and vice versa)

Definitions related to entropy and information

- Chain rule

$$H(X,Y) = H(X) + H(Y\,|\,X) = H(Y) + H(X\,|\,Y)$$

- Mutual information

$$\begin{aligned} I(X;Y) &= H(X) - H(X\,|\,Y) = H(Y) - H(Y\,|\,X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

  - $I(X;Y) = I(Y;X)$
  - average reduction in uncertainty about X when learning value of Y (and vice versa)

Definitions related to entropy and information

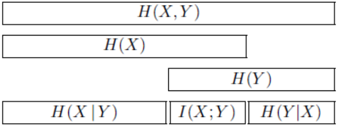
- Chain rule

$$H(X,Y) = H(X) + H(Y \mid X) = H(Y) + H(X \mid Y)$$

- Mutual information

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

  - $I(X;Y) = I(Y;X)$
  - average reduction in uncertainty about X when learning value of Y (and vice versa)



(D. MacKay, Information Theory, Inference, and Learning Algorithms)

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x,y) = p(x)p(y)$.
- $H(X,Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X \mid Y) = H(X)$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x, y) = p(x)p(y)$.
- $H(X, Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X \mid Y) = H(X)$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x,y) = p(x)p(y)$.
- $H(X,Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

$$= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log P(X=x) + P(X=x)P(Y=y)\log P(Y=y$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X \mid Y) = H(X)$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x, y) = p(x)p(y)$.
- $H(X, Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

$$= -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P(X=x)P(Y=y)\log P(X=x) + P(X=x)P(Y=y)\log P(Y=y$$

$$= -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x) - \sum_{y\in\mathcal{Y}} P(Y=y)\log P(Y=y)$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X \mid Y) = H(X)$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x, y) = p(x)p(y)$.
- $H(X, Y) = H(X) + H(Y)$

$$H(X, Y) = -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x)P(Y = y) \log(P(X = x)P(Y = y))$$

$$= -\sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P(X = x)P(Y = y) \log P(X = x) + P(X = x)P(Y = y) \log P(Y = y$$

$$= -\sum_{x \in \mathcal{X}} P(X = x) \log P(X = x) - \sum_{y \in \mathcal{Y}} P(Y = y) \log P(Y = y)$$

- $I(X; Y) = H(X) + H(Y) - \underbrace{H(X, Y)}_{H(X)+H(Y)} = 0$
- $H(X \mid Y) = H(X)$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x,y) = p(x)p(y)$.
- $H(X,Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

$$= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log P(X=x) + P(X=x)P(Y=y)\log P(Y=y$$

$$= -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x) - \sum_{y\in\mathcal{Y}} P(Y=y)\log P(Y=y)$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X\,|\,Y) = H(X)$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x, y) = p(x)p(y)$.
- $H(X, Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

$$= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log P(X=x) + P(X=x)P(Y=y)\log P(Y=y$$

$$= -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x) - \sum_{y\in\mathcal{Y}} P(Y=y)\log P(Y=y)$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X\,|\,Y) = H(X)$

$$H(X\,|\,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log \underbrace{P(X=x\,|\,Y=y)}_{P(X=x)}$$

Definitions related to entropy and information

Independence of X and Y

- Under independence of X and Y, $p(x,y) = p(x)p(y)$.
- $H(X,Y) = H(X) + H(Y)$

$$H(X,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log(P(X=x)P(Y=y))$$

$$= -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log P(X=x) + P(X=x)P(Y=y)\log P(Y=y$$

$$= -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x) - \sum_{y\in\mathcal{Y}} P(Y=y)\log P(Y=y)$$

- $I(X;Y) = H(X) + H(Y) - \underbrace{H(X,Y)}_{H(X)+H(Y)} = 0$
- $H(X\,|\,Y) = H(X)$

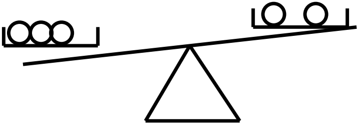
$$H(X\,|\,Y) = -\sum_{x\in\mathcal{X},y\in\mathcal{Y}} P(X=x)P(Y=y)\log \underbrace{P(X=x\,|\,Y=y)}_{P(X=x)}$$

$$= -\sum_{x\in\mathcal{X}} P(X=x)\log P(X=x)$$

Entropy in action

The optimal weighing problem

- Given 12 balls, all equal except for one that is lighter or heavier.
- What is the ideal weighting strategy and how many weighings are needed to identify the odd ball **and** tell if it is lighter or heavier?



Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(X = x) \log \frac{P(X = x)}{Q(Y = y)}$$

- $D_{KL}(P, \|Q) \neq D_{KL}(Q\|\|P)$ (not symmetric)
- $D_{KL}(P, \|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(X=x) \log \frac{P(X=x)}{Q(Y=y)}$$

- $D_{KL}(P, \|Q) \neq D_{KL}(Q\|\|P)$
(not symmetric)
- $D_{KL}(P, \|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(X=x) \log \frac{P(X=x)}{Q(Y=y)}$$

- $D_{KL}(P, \|Q) \neq D_{KL}(Q\|\|P)$
  (not symmetric)
- $D_{KL}(P, \|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

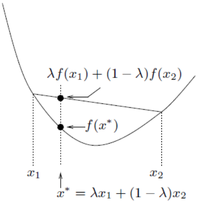
$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(X=x)\log\frac{P(X=x)}{Q(Y=y)}$$

- $D_{KL}(P,\|Q) \neq D_{KL}(Q\|\|P)$ (not symmetric)
- $D_{KL}(P,\|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Definition of convexity:

- $f(x)$ is convex over interval (a,b), if $\forall x_1, x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$

$$\begin{aligned} & f(\lambda x_1 + (1-\lambda)x_2) \\ & \leq \lambda f(x_1) + (1-\lambda)f(x_2) \end{aligned}$$



(D. MacKay, Information Theory, Inference, and Learning Algorithms)

Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

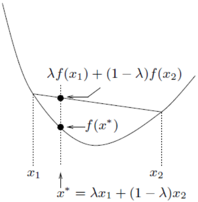
$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(X = x) \log \frac{P(X = x)}{Q(Y = y)}$$

- $D_{KL}(P, \|Q) \neq D_{KL}(Q\|\|P)$ (not symmetric)
- $D_{KL}(P, \|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Definition of convexity:

- $f(x)$ is convex over interval (a, b), if $\forall x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$

$$\begin{aligned} & f(\lambda x_1 + (1 - \lambda) x_2) \\ \leq \; & \lambda f(x_1) + (1 - \lambda) f(x_2) \end{aligned}$$



(D. MacKay, Information Theory, Inference, and Learning Algorithms)

Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

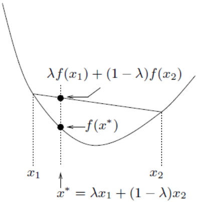
$$D_{KL}(P\|Q) = \sum_{x\in\mathcal{X}} P(X=x)\log\frac{P(X=x)}{Q(Y=y)}$$

- $D_{KL}(P,\|Q) \neq D_{KL}(Q\|\|P)$ (not symmetric)
- $D_{KL}(P,\|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Definition of convexity:

- $f(x)$ is convex over interval (a,b), if $\forall x_1, x_2 \in (a,b)$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \\ \leq \lambda f(x_1) + (1-\lambda)f(x_2)$$



(D. MacKay, Information Theory, Inference, and Learning Algorithms)

Kulback-Leibler divergence

- For two probability distributions over X, $P(X)$ and $Q(X)$, the KL divergence (or relative Entropy) is defined as

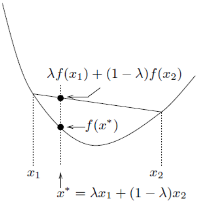
$$D_{KL}(P\|Q) = \sum_{x \in \mathcal{X}} P(X=x) \log \frac{P(X=x)}{Q(Y=y)}$$

- $D_{KL}(P, \|Q) \neq D_{KL}(Q\|\|P)$ (not symmetric)
- $D_{KL}(P, \|Q)$ is strictly convex.
- $D_{KL}(P\|Q) \geq 0$ (Gibb's inequality)
- $D_{KL}(P\|Q) = 0$ if and only if $P = Q$.
- KL divergence will be useful as scoring function for approximations Q of probability distributions P that are intractable.

Definition of convexity:

- $f(x)$ is convex over interval (a, b), if $\forall x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

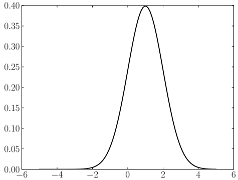


(D. MacKay, Information Theory, Inference, and Learning Algorithms)

Probability distributions

- Normal distribution (Gaussian distribution)

$$p(x \mid \mu, \sigma^2) = \mathcal{N}\left(\, x \mid \mu, \sigma^2 \,\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



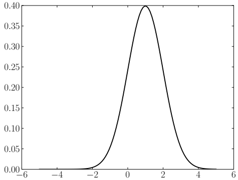
- Multivariate normal distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\, \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \,\right)$$

Probability distributions

- Normal distribution (Gaussian distribution)

$$p(x \mid \mu, \sigma^2) = \mathcal{N}\left(x \mid \mu, \sigma^2 \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Multivariate normal distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$$

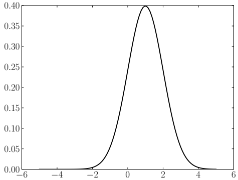
$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu}) \right]$$

- data term

Probability distributions

- Normal distribution (Gaussian distribution)

$$p(x \mid \mu, \sigma^2) = \mathcal{N}\left(\, x \mid \mu, \sigma^2 \,\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Multivariate normal distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\, \boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \,\right)$$

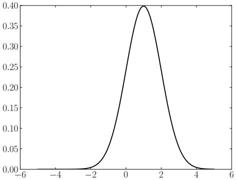
$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]$$

- data term normalization constant

Probability distributions

- Normal distribution (Gaussian distribution)

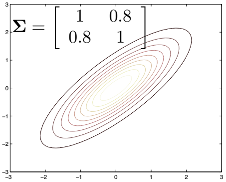
$$p(x \mid \mu, \sigma^2) = \mathcal{N}\left(x \mid \mu, \sigma^2 \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Multivariate normal distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$$

$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right]$$

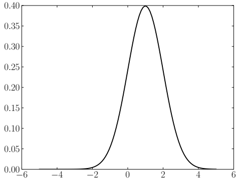


- data term normalization constant

Probability distributions

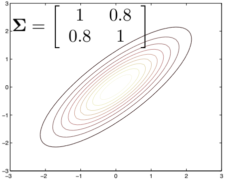
- Normal distribution (Gaussian distribution)

$$p(x \mid \mu, \sigma^2) = \mathcal{N}\left(x \mid \mu, \sigma^2 \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Multivariate normal distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$$
$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right]$$

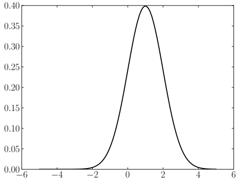


- data term normalization constant

Probability distributions

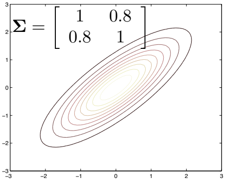
- Normal distribution (Gaussian distribution)

$$p(x \mid \mu, \sigma^2) = \mathcal{N}\left(x \mid \mu, \sigma^2 \right) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$



- Multivariate normal distribution

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma} \right)$$
$$= \frac{1}{\sqrt{|2\pi\boldsymbol{\Sigma}|}} \cdot \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \right]$$



- data term   normalization constant

Probability distributions

continued...

- Bernoulli

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

- Gamma

$$p(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

Probability distributions

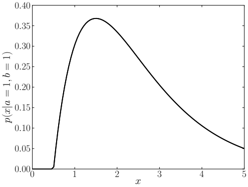
continued...

- Bernoulli

$$p(x \mid \theta) = \theta^x (1-\theta)^{1-x}$$

- Gamma

$$p(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$



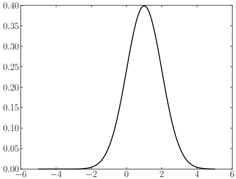
Probability distributions

The Gaussian revisited

- Gaussian PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Positive: $\mathcal{N}\left(x \mid \mu, \sigma^2\right) > 0$
- Normalized: $\int_{-\infty}^{+\infty} \mathcal{N}\left(x \mid \mu, \sigma\right) \mathrm{d}x = 1$ (check)
- Expectation:
  $< x >= \int_{-\infty}^{+\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x \mathrm{d}x = \mu$
- Variance: $\mathrm{Var}[x] =< x^2 > - < x >^2$
  $= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$



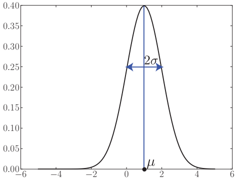
Probability distributions

The Gaussian revisited

- Gaussian PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Positive: $\mathcal{N}\left(x \mid \mu, \sigma^2\right) > 0$
- Normalized: $\int_{-\infty}^{+\infty} \mathcal{N}\left(x \mid \mu, \sigma\right) \mathrm{d}x = 1$ (check)
- Expectation:
  $< x >= \int_{-\infty}^{+\infty} \mathcal{N}\left(x \mid \mu, \sigma^2\right) x \mathrm{d}x = \mu$
- Variance: $\mathrm{Var}[x] =< x^2 > - < x >^2$
  $= \mu^2 + \sigma^2 - \mu^2 = \sigma^2$



Inference for the normal distribution

Ingredients

- Data sampled from unknown distribution $p(\mathcal{D} \,|\, \boldsymbol{\theta}_0)$

$$\mathcal{D} = \{x_1, \ldots, x_N\} \sim p(\mathcal{D} \,|\, \boldsymbol{\theta}_0)$$

- Model $\mathcal{H}_{\text{Gauss}}$ – normal PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$
$$\boldsymbol{\theta} = \{\mu, \sigma^2\}$$

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

Inference for the normal distribution

Ingredients

- Data sampled from unknown distribution $p(\mathcal{D} \mid \boldsymbol{\theta}_0)$

$$\mathcal{D} = \{x_1, \ldots, x_N\} \sim p(\mathcal{D} \mid \boldsymbol{\theta}_0)$$

- Model $\mathcal{H}_{Gauss}$ – normal PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\boldsymbol{\theta} = \{\mu, \sigma^2\}$$

- Likelihood

$$p(\mathcal{D} \mid \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

Inference for the normal distribution

Ingredients

- Data sampled from unknown distribution $p(\mathcal{D} \,|\, \boldsymbol{\theta}_0)$

$$\mathcal{D} = \{x_1, \ldots, x_N\} \sim p(\mathcal{D} \,|\, \boldsymbol{\theta}_0)$$

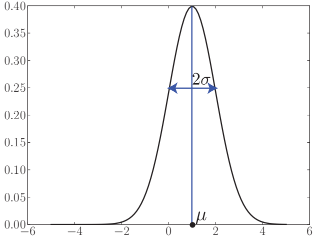
- Model $\mathcal{H}_{\mathsf{Gauss}}$ – normal PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\boldsymbol{\theta} = \{\mu, \sigma^2\}$$

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$



Inference for the normal distribution

Ingredients

- Data sampled from unknown distribution $p(\mathcal{D} \,|\, \boldsymbol{\theta}_0)$

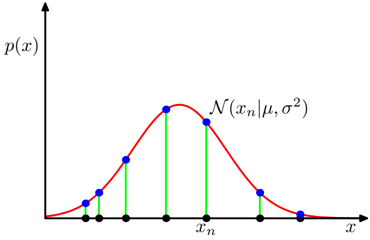
$$\mathcal{D} = \{x_1, \ldots, x_N\} \sim p(\mathcal{D} \,|\, \boldsymbol{\theta}_0)$$

- Model $\mathcal{H}_{\mathsf{Gauss}}$ – normal PDF

$$\mathcal{N}\left(\, x \mid \mu, \sigma^2 \,\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$
$$\boldsymbol{\theta} = \{\mu, \sigma^2\}$$

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(\, x_n \mid \mu, \sigma^2 \,\right)$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Inference for the normal distribution

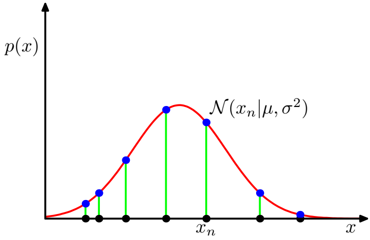
Maximum likelihood

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

- Maximum likelihood
- Chose parameters $\hat{\mu}$ and $\hat{\sigma}^2$ that *maximize* the likelihood of $\mathcal{D}$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D} \,|\, \boldsymbol{\theta})$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Inference for the normal distribution

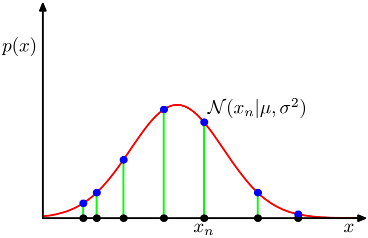
Maximum likelihood

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

- Maximum likelihood
- Chose parameters $\hat{\mu}$ and $\hat{\sigma}^2$ that *maximize* the likelihood of $\mathcal{D}$

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, p(\mathcal{D} \,|\, \boldsymbol{\theta})$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Inference for the normal distribution

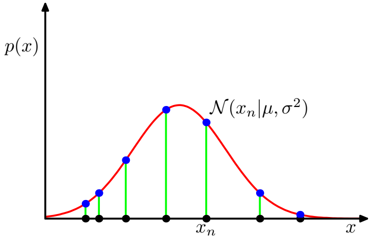
Maximum likelihood

- Likelihood

$$p(\mathcal{D} \,|\, \boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

- Maximum likelihood
- Chose parameters $\hat{\mu}$ and $\hat{\sigma}^2$ that *maximize* the likelihood of $\mathcal{D}$

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} p(\mathcal{D} \,|\, \boldsymbol{\theta})$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Maximum likelihood estimation in the normal distribution

- Data sample $\mathcal{D}$ of size N modeled by a univariate normal distribution
- Likelihood of the data under the model $p(\mathcal{D} \mid \mu, \sigma^2)$:

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

- Equivalently maximize the *log-Likelihood* $\log p(\mathcal{D} \mid \mu, \sigma^2) = \mathcal{L}\left(\mu, \sigma^2\right)$

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

$$= \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

Maximum likelihood estimation in the normal distribution

- Data sample $\mathcal{D}$ of size N modeled by a univariate normal distribution
- Likelihood of the data under the model $p(\mathcal{D} \mid \mu, \sigma^2)$:

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

- Equivalently maximize the *log-Likelihood* $\log p(\mathcal{D} \mid \mu, \sigma^2) = \mathcal{L}\left(\mu, \sigma^2\right)$

Maximum likelihood estimation in the normal distribution

- Data sample $\mathcal{D}$ of size N modeled by a univariate normal distribution
- Likelihood of the data under the model $p(\mathcal{D} \mid \mu, \sigma^2)$:

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Equivalently maximize the *log-Likelihood* $\log p(\mathcal{D} \mid \mu, \sigma^2) = \mathcal{L}\left(\mu, \sigma^2\right)$

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \sum_{n=1}^{N} -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

Maximum likelihood estimation in the normal distribution

- Data sample $\mathcal{D}$ of size N modeled by a univariate normal distribution
- Likelihood of the data under the model $p(\mathcal{D} \mid \mu, \sigma^2)$:

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Equivalently maximize the *log-Likelihood* $\log p(\mathcal{D} \mid \mu, \sigma^2) = \mathcal{L}\left(\mu, \sigma^2\right)$

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

Maximum likelihood estimation in the normal distribution

- Data sample $\mathcal{D}$ of size N modeled by a univariate normal distribution
- Likelihood of the data under the model $p(\mathcal{D} \mid \mu, \sigma^2)$:

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Equivalently maximize the *log-Likelihood* $\log p(\mathcal{D} \mid \mu, \sigma^2) = \mathcal{L}\left(\mu, \sigma^2\right)$

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

Maximum likelihood estimation in the normal distribution

- Data sample $\mathcal{D}$ of size N modeled by a univariate normal distribution
- Likelihood of the data under the model $p(\mathcal{D} \mid \mu, \sigma^2)$:

$$\prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- Equivalently maximize the *log-Likelihood* $\log p(\mathcal{D} \mid \mu, \sigma^2) = \mathcal{L}\left(\mu, \sigma^2\right)$

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

$$= \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} =$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} =$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} =$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} =$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \hat{\mu})^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n - \hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n\right) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} =$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n\right) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n - \hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}(\sum_{n=1}^{N} x_n) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n\right) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n - \hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}\left(\sum_{n=1}^{N} x_n\right) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n - \hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}(\sum_{n=1}^{N} x_n) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n - \hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu,\sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n-\mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu,\sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu,\sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n-\hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}\left(\sum_{n=1}^{N}x_n\right) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu},\sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu},\sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n-\mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n-\hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n-\hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n-\hat{\mu})^2$$

Maximum likelihood estimation in the normal distribution

$$\mathcal{L}\left(\mu, \sigma^2\right) = \sum_{n=1}^{N} -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x_n - \mu)^2$$

- Take the derivative of $\mathcal{L}\left(\mu, \sigma^2\right)$ with respect to μ:

$$\frac{\partial \mathcal{L}\left(\mu, \sigma^2\right)}{\partial \mu} = -\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \mu)$$

- set to zero and solve for $\hat{\mu}$:

$$-\frac{1}{\sigma^2}\sum_{n=1}^{N}(x_n - \hat{\mu}) = 0$$

$$-\frac{1}{\sigma^2}(\sum_{n=1}^{N} x_n) + \frac{N}{\sigma^2}\hat{\mu} = 0$$

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n \quad \text{sample mean}$$

- Take the derivative of $\mathcal{L}\left(\hat{\mu}, \sigma^2\right)$ with respect to σ^2:

$$\frac{\partial \mathcal{L}\left(\hat{\mu}, \sigma^2\right)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \sum_{n=1}^{N}\frac{1}{2\sigma^4}(x_n - \mu)^2$$

- set to zero and solve for $\hat{\sigma}^2$:

$$-\frac{N}{2\hat{\sigma}^2} + \sum_{n=1}^{N}\frac{1}{2\hat{\sigma}^4}(x_n - \hat{\mu})^2 = 0$$

$$\frac{N\hat{\sigma}^2}{2} = \sum_{n=1}^{N}\frac{1}{2}(x_n - \hat{\mu})^2$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2 \quad \text{sample variance}$$

Inference for the Gaussian

Maximum likelihood

- Maximum likelihood solutions

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

  Equivalent to common mean and variance estimators (almost).

- Maximum likelihood ignores parameter uncertainty
  - Think of the ML solution for a single observed datapoint x_1

$$\hat{\mu} = x_1$$

$$\hat{\sigma}^2 = (x_1 - \hat{\mu})^2 = 0$$

- How about **Bayesian inference**?

Inference for the Gaussian

Maximum likelihood

- Maximum likelihood solutions

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

  Equivalent to common mean and variance estimators (almost).
- Maximum likelihood ignores parameter uncertainty
  - Think of the ML solution for a single observed datapoint x_1

$$\hat{\mu} = x_1$$
$$\hat{\sigma}^2 = (x_1 - \hat{\mu})^2 = 0$$

- How about **Bayesian inference**?

Inference for the Gaussian

Maximum likelihood

- Maximum likelihood solutions

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N} (x_n - \hat{\mu})^2$$

  Equivalent to common mean and variance estimators (almost).
- Maximum likelihood ignores parameter uncertainty
  - Think of the ML solution for a single observed datapoint x_1

$$\hat{\mu} = x_1$$

$$\hat{\sigma}^2 = (x_1 - \hat{\mu})^2 = 0$$

- How about **Bayesian inference**?

Outline

The Rules of Probability

Sum & Product Rule

Sum Rule	$p(x) = \sum_y p(x, y)$
Product Rule	$p(x, y) = p(y \mid x) p(x)$

Bayes Theorem

- Using the product rule we obtain

$$p(y \mid x) = \frac{p(x \mid y) p(y)}{p(x)}$$

$$p(x) = \sum_y p(x \mid y) p(y)$$

The Rules of Probability

Sum & Product Rule

$$\begin{array}{cc} \text{Sum Rule} & p(x) = \sum_y p(x, y) \\ \text{Product Rule} & p(x, y) = p(y \mid x) p(x) \end{array}$$

Bayes Theorem

- Using the product rule we obtain

$$p(y \mid x) = \frac{p(x \mid y) p(y)}{p(x)}$$

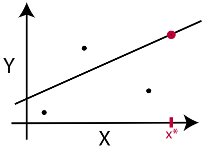
$$p(x) = \sum_y p(x \mid y) p(y)$$

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.

- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



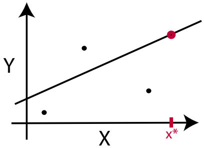
- In maximum likelihood estimation we maximized $p(\mathcal{D} \mid \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters θ given Data $\mathcal{D}$ using Bayes theorem.

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \,|\, \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat θ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters θ given Data $\mathcal{D}$ using Bayes theorem.

$$= \frac{\boxed{p(\mathcal{D} \,|\, \boldsymbol{\theta})}}{}$$

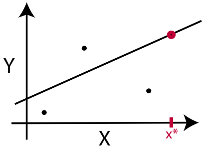
- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.

- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \mid \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters θ given Data $\mathcal{D}$ using Bayes theorem.

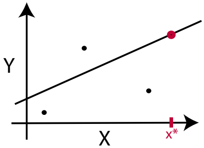
$$= \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{}$$

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \mid \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$ using Bayes theorem.

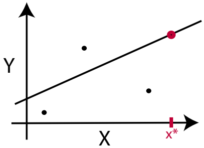
$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{}$$

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \mid \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$ using Bayes theorem.

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{}$$

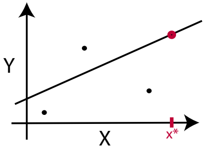
posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \,|\, \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$ using Bayes theorem.

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

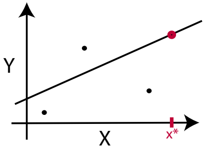
posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.
- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \mid \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$ using Bayes theorem.

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D}\,,\ \boldsymbol{\theta})\ \mathrm{d}\boldsymbol{\theta}}$$

posterior $\propto$ likelihood $\cdot$ prior

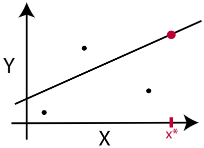
- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

Bayesian probability calculus

- Bayes rule is the basis for Bayesian inference and learning.

- Assume we have a model with parameters $\boldsymbol{\theta}$, e.g.

$$y = \theta_0 + \theta_1 \cdot x + \epsilon$$



- In maximum likelihood estimation we maximized $p(\mathcal{D} \,|\, \boldsymbol{\theta})$ w.r.t $\boldsymbol{\theta}$
- Idea: treat $\boldsymbol{\theta}$ as a random variable under $p(\boldsymbol{\theta})$
- Infer the conditional distribution of the parameters $\boldsymbol{\theta}$ given Data $\mathcal{D}$ using Bayes theorem.

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) \; \mathrm{d}\boldsymbol{\theta}}$$

posterior $\propto$ likelihood $\cdot$ prior

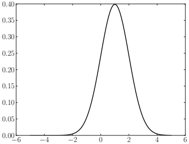
- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ I

- Likelihood:

$$p(\mathcal{D} \,|\, \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$



- Specify normal prior on μ:

- $p(\mathcal{D})$ not needed for MAP estimation (constant in the parameter).

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

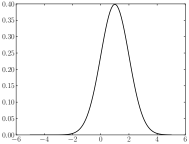
“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ I

- Likelihood:

$$p(\mathcal{D} \mid \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

- Specify normal prior on μ:



- $p(\mathcal{D})$ not needed for MAP estimation (constant in the parameter).

$$\boxed{p(\boldsymbol{\theta} \mid \mathcal{D})} = \frac{\boxed{p(\mathcal{D} \mid \boldsymbol{\theta})} \cdot \boxed{p(\boldsymbol{\theta})}}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

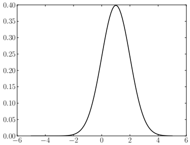
“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ I

- Likelihood:

$$p(\mathcal{D} \,|\, \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2 \right)$$

- Specify normal prior on μ: $p(\mu) = \mathcal{N}\left(\mu \mid m_0, s_0^2 \right)$



- $p(\mathcal{D})$ not needed for MAP estimation (constant in the parameter).

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

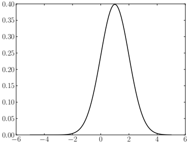
"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ I

- Likelihood:

$$p(\mathcal{D} \,|\, \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(\, x_n \mid \mu, \sigma^2 \,\right)$$

- Specify normal prior on μ: $p(\mu) = \mathcal{N}\left(\, \mu \mid m_0, s_0^2 \,\right)$



$$\boxed{p(\boldsymbol{\theta} \,|\, \mathcal{D})} \propto \boxed{\prod_{n=1}^{N} \mathcal{N}\left(\, x_n \mid \mu, \sigma^2 \,\right)} \cdot \boxed{\mathcal{N}\left(\, \mu \mid m_0, s_0^2 \,\right)}$$

- $p(\mathcal{D})$ not needed for MAP estimation (constant in the parameter).

$$\boxed{p(\boldsymbol{\theta} \,|\, \mathcal{D})} = \frac{\boxed{p(\mathcal{D} \,|\, \boldsymbol{\theta})} \cdot \boxed{p(\boldsymbol{\theta})}}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

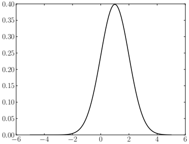
- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ I

- Likelihood:

$$p(\mathcal{D} \,|\, \mu, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(\, x_n \mid \mu, \sigma^2 \,\right)$$



- Specify normal prior on μ: $\;p(\mu) = \mathcal{N}\left(\, \mu \mid m_0, s_0^2 \,\right)$

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) \propto \prod_{n=1}^{N} \mathcal{N}\left(\, x_n \mid \mu, \sigma^2 \,\right) \cdot \mathcal{N}\left(\, \mu \mid m_0, s_0^2 \,\right)$$

- $p(\mathcal{D})$ not needed for MAP estimation (constant in the parameter).

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) \cdot \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

- take logarithm of the posterior

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}) = Z + \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) + \log \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) \cdot \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

- take logarithm of the posterior

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}) = Z + \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) + \log \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$p(\boldsymbol{\theta} \mid \mathcal{D}) \propto \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) \cdot \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

- take logarithm of the posterior

$$\log p(\boldsymbol{\theta} \mid \mathcal{D}) = Z + \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) + \log \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

$$= Z + -\frac{1}{2}\left(\sum_{n=1}^{N} \log(2\pi\sigma^2) + \frac{1}{\sigma^2}(x_n - \mu)^2\right) - \frac{1}{2}\left(\log(2\pi\sigma_\mu^2) + \frac{1}{s_0^2}(\mu - m_0)^2\right)$$

$$p(\boldsymbol{\theta} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood $\cdot$ prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) \propto \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) \cdot \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

- take logarithm of the posterior

$$\log p(\boldsymbol{\theta} \,|\, \mathcal{D}) = Z + \sum_{n=1}^{N} \log \mathcal{N}\left(x_n \mid \mu, \sigma^2\right) + \log \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

$$= Z + -\frac{1}{2}\left(\sum_{n=1}^{N} \log(2\pi\sigma^2) + \frac{1}{\sigma^2}(x_n - \mu)^2\right) - \frac{1}{2}\left(\log(2\pi\sigma_\mu^2) + \frac{1}{s_0^2}(\mu - m_0)^2\right)$$

$$= Z' - -\frac{1}{2}\left(\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)^2\right) - \frac{1}{s_0^2}(\mu - m_0)^2\right)$$

$$p(\boldsymbol{\theta} \,|\, \mathcal{D}) = \frac{p(\mathcal{D} \,|\, \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

posterior $\propto$ likelihood · prior

- Likelihood
- Prior
- Posterior
- Marginal likelihood (normalization constant)

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \,|\, \mathcal{D})} = Z' - \boxed{-\tfrac{1}{2}\left(\left(\sum_{n=1}^{N} \tfrac{1}{\sigma^2}(x_n - \mu)^2\right) - \tfrac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \,|\, \mathcal{D})}{\partial \mu} =$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - **shrinkage** parameter
  - **regularization** parameter

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \mid \mathcal{D})} = Z' - \boxed{-\tfrac{1}{2}\left(\left(\sum_{n=1}^{N} \tfrac{1}{\sigma^2}(x_n - \mu)^2\right) - \tfrac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \mid \mathcal{D})}{\partial \mu} =$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - shrinkage parameter
  - regularization parameter

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \,|\, \mathcal{D})} = Z' - \boxed{-\tfrac{1}{2}\left(\left(\sum_{n=1}^{N} \tfrac{1}{\sigma^2}(x_n - \mu)^2\right) - \tfrac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \,|\, \mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

- where $\delta = \frac{\sigma^2}{s_0}$
  - shrinkage parameter
  - regularization parameter

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \,|\, \mathcal{D})} = Z' - \boxed{-\tfrac{1}{2}\left(\left(\textstyle\sum_{n=1}^{N} \tfrac{1}{\sigma^2}(x_n - \mu)^2\right) - \tfrac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \,|\, \mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - **shrinkage** parameter
  - **regularization** parameter

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \,|\, \mathcal{D})} = Z' - \boxed{-\frac{1}{2}\left(\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)^2\right) - \frac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \,|\, \mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- where $\delta = \frac{\sigma^2}{s_0}$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

$$\left(\frac{N}{\sigma^2} - \frac{1}{s_0^2}\right)\mu_{MAP} = \frac{1}{s_0^2} m_0 + \sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n)$$

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\log p(\boldsymbol{\theta} \,|\, \mathcal{D}) = Z' - -\frac{1}{2}\left(\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)^2\right) - \frac{1}{s_0^2}(\mu - m_0)^2\right)$$

- Take derivative

$$\frac{\partial p(\mu \,|\, \mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

$$\left(\frac{N}{\sigma^2} - \frac{1}{s_0^2}\right)\mu_{MAP} = \frac{1}{s_0^2} m_0 + \sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n)$$

$$\mu_{MAP} = \frac{\delta}{N - \delta} m_0 + \frac{1}{N - \delta} \sum_{n=1}^{N} x_n$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - shrinkage parameter
  - regularization parameter

"Bayesian estimation" in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta}\,|\,\mathcal{D})} = Z' - \boxed{-\frac{1}{2}\left(\left(\sum_{n=1}^{N}\frac{1}{\sigma^2}(x_n-\mu)^2\right) - \frac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu\,|\,\mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N}\frac{1}{\sigma^2}(x_n-\mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N}\frac{1}{\sigma^2}(x_n-\mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

$$\left(\frac{N}{\sigma^2} - \frac{1}{s_0^2}\right)\mu_{MAP} = \frac{1}{s_0^2}m_0 + \sum_{n=1}^{N}\frac{1}{\sigma^2}(x_n)$$

$$\mu_{MAP} = \frac{\delta}{N-\delta}m_0 + \frac{1}{N-\delta}\sum_{n=1}^{N}x_n$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - **shrinkage** parameter
  - **regularization** parameter

“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \mid \mathcal{D})} = Z' - \boxed{-\frac{1}{2}\left(\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)^2\right) - \frac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \mid \mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

$$\left(\frac{N}{\sigma^2} - \frac{1}{s_0^2}\right)\mu_{MAP} = \frac{1}{s_0^2}m_0 + \sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n)$$

$$\mu_{MAP} = \frac{\delta}{N - \delta}m_0 + \frac{1}{N - \delta}\sum_{n=1}^{N} x_n$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - shrinkage parameter
  - regularization parameter

$$\mu_{MAP} = \frac{\delta}{N - \delta}m_0 + \frac{N}{N - \delta}\hat{\mu}$$

“Bayesian estimation” in the normal distribution

Maximum a posteriori estimation of the mean μ II

$$\boxed{\log p(\boldsymbol{\theta} \,|\, \mathcal{D})} = Z' - \boxed{-\tfrac{1}{2}\left(\left(\sum_{n=1}^{N} \tfrac{1}{\sigma^2}(x_n - \mu)^2\right) - \tfrac{1}{s_0^2}(\mu - m_0)^2\right)}$$

- Take derivative

$$\frac{\partial p(\mu \,|\, \mathcal{D})}{\partial \mu} = -\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu)\right) - \frac{1}{s_0^2}(\mu - m_0)$$

- set to zero and solve for μ_{MAP}

$$-\left(\sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n - \mu_{MAP})\right) - \frac{1}{s_0^2}(\mu_{MAP} - m_0) = 0$$

$$\left(\frac{N}{\sigma^2} - \frac{1}{s_0^2}\right)\mu_{MAP} = \frac{1}{s_0^2}m_0 + \sum_{n=1}^{N} \frac{1}{\sigma^2}(x_n)$$

$$\mu_{MAP} = \frac{\delta}{N-\delta}m_0 + \frac{1}{N-\delta}\sum_{n=1}^{N} x_n$$

- where $\delta = \dfrac{\sigma^2}{s_0}$
  - **shrinkage** parameter
  - **regularization** parameter

$$\mu_{MAP} = \frac{\delta}{N-\delta}m_0 + \frac{N}{N-\delta}\hat{\mu}$$

Bayesian Inference for the Gaussian

Ingredients

- Data

$$\mathcal{D} = \{x_1, \ldots, x_N\}$$

- Model $\mathcal{H}_{Gauss}$ – Gaussian PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\boldsymbol{\theta} = \{\mu\}$$

  - For simplicity: assume variance σ^2 is known.

- Likelihood

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

Bayesian Inference for the Gaussian

Ingredients

- Data

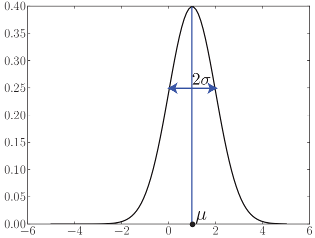
$$\mathcal{D} = \{x_1, \ldots, x_N\}$$

- Model $\mathcal{H}_{Gauss}$ – Gaussian PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

$$\boldsymbol{\theta} = \{\mu\}$$

  - For simplicity: assume variance σ^2 is known.



- Likelihood

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$

Bayesian Inference for the Gaussian

Ingredients

- Data

$$\mathcal{D} = \{x_1, \ldots, x_N\}$$

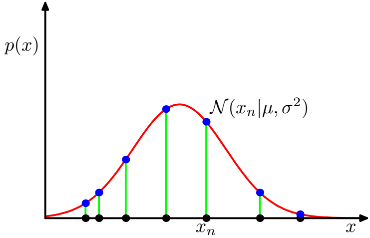
- Model $\mathcal{H}_{\mathsf{Gauss}}$ – Gaussian PDF

$$\mathcal{N}\left(x \mid \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$
$$\boldsymbol{\theta} = \{\mu\}$$

  - For simplicity: assume variance σ^2 is known.

- Likelihood

$$p(\mathcal{D} \mid \mu) = \prod_{n=1}^{N} \mathcal{N}\left(x_n \mid \mu, \sigma^2\right)$$



(C.M. Bishop, Pattern Recognition and Machine Learning)

Bayesian Inference for the Gaussian

Bayes rule

- Combine likelihood with a Gaussian prior over μ

$$p(\mu) = \mathcal{N}\left(\mu \mid m_0, s_0^2\right)$$

- The posterior is proportional to

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) p(\mu)$$

Bayesian Inference for the Gaussian

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) \cdot p(\mu)$$
$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu - m_0)^2}$$

- Posterior parameters follow as the new coefficients.
- Note: Posterior has form of normal distribution, thus is normalized

Bayesian Inference for the Gaussian

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) \cdot p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2 s_0^2}(\mu - m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2 s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2\mu x_n + x_n^2) \right]$$

- Posterior parameters follow as the new coefficients.
- Note: Posterior has form of normal distribution, thus is normalized

Bayesian Inference for the Gaussian

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) \cdot p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu - m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2\mu x_n + x_n^2) \right]$$

$$= C2 \exp\left[-\frac{1}{2} \underbrace{\left(\frac{1}{s_0^2} + \frac{N}{\sigma^2} \right)}_{1/s_P^2} \left(\mu^2 - 2\mu \underbrace{\hat{\sigma}\left(\frac{1}{s_0^2} m_0 + \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n \right)}_{m_P} \right) + C3 \right]$$

- Posterior parameters follow as the new coefficients.
- Note: Posterior has form of normal distribution, thus is normalized

Bayesian Inference for the Gaussian

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) \cdot p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n-\mu)^2}\right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu-m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(\mu^2 - 2\mu x_n + x_n^2)\right]$$

$$= C2 \exp\left[-\frac{1}{2}\underbrace{\left(\frac{1}{s_0^2} + \frac{N}{\sigma^2}\right)}_{1/s_P^2}\left(\mu^2 - 2\mu \underbrace{\hat{\sigma}\left(\frac{1}{s_0^2} m_0 + \frac{1}{\sigma^2}\sum_{n=1}^{N} x_n\right)}_{m_P}\right) + C3\right]$$

- Posterior parameters follow as the new coefficients.
- Note: Posterior has form of normal distribution, thus is normalized

Bayesian Inference for the Gaussian

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) \cdot p(\mu)$$

$$= \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x_n - \mu)^2} \right] \frac{1}{\sqrt{2\pi s_0^2}} e^{-\frac{1}{2s_0^2}(\mu - m_0)^2}$$

$$= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}^{N} \frac{1}{\sqrt{2\pi s_0^2}}}_{C1} \exp\left[-\frac{1}{2s_0^2}(\mu^2 - 2\mu m_0 + m_0^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2\mu x_n + x_n^2) \right]$$

$$= C2 \exp\left[-\frac{1}{2} \underbrace{\left(\frac{1}{s_0^2} + \frac{N}{\sigma^2} \right)}_{1/s_P^2} \left(\mu^2 - 2\mu \underbrace{\hat{\sigma}\left(\frac{1}{s_0^2} m_0 + \frac{1}{\sigma^2} \sum_{n=1}^{N} x_n\right)}_{m_P} \right) + C3 \right]$$

- Posterior parameters follow as the new coefficients.
- Note: Posterior has form of normal distribution, thus is normalized

Bayesian Inference for the Gaussian

- Posterior of the mean: $p(\mu \,|\, \mathcal{D}, \sigma^2) \propto \mathcal{N}\left(\mu \mid m_P, s_P \right)$, after some rewriting

$$m_P = \frac{\sigma^2}{Ns_0^2 + \sigma^2} m_0 + \frac{Ns_0^2}{Ns_0^2 + \sigma^2} \hat{\mu}, \qquad \hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

$$\frac{1}{s_P^2} = \frac{1}{s_0^2} + \frac{N}{\sigma^2}$$

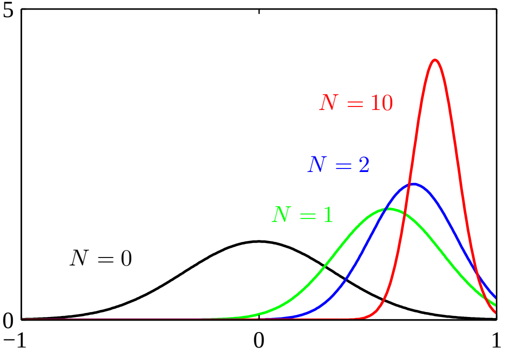
- Limiting cases for no and infinite amount of data

	$N = 0$	$N \rightarrow \infty$
m_P	m_0	$\hat{\mu}$
s_P^2	s_0^2	0

Bayesian Inference for the Gaussian

Examples

- Posterior $p(\mu \,|\, \mathcal{D}, \sigma^2)$ for increasing data sizes.



(C.M. Bishop, Pattern Recognition and Machine Learning)

Conjugate priors

- It is not chance that the posterior

$$p(\mu \mid \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \mid \mu, \sigma^2) p(\mu)$$

is tractable in closed form for the Gaussian.

Conjugate prior

$p(\theta)$ is a conjugate prior for a particular likelihood $p(\mathcal{D} \mid \theta)$ if the posterior is of the same functional form than the prior.

Conjugate priors

- It is not chance that the posterior

$$p(\mu \,|\, \mathcal{D}, \sigma^2) \propto p(\mathcal{D} \,|\, \mu, \sigma^2) p(\mu)$$

is tractable in closed form for the Gaussian.

Conjugate prior

$p(\theta)$ is a conjugate prior for a particular likelihood $p(\mathcal{D} \,|\, \theta)$ if the posterior is of the same functional form than the prior.

Conjugate priors

Exponential family distributions

- A large class of probability distributions are part of the exponential family (all in this course) and can be written as:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x})g(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^\top \boldsymbol{u}(\boldsymbol{x})\}$$

- For example for the Gaussian:

$$\begin{aligned} p(x \mid \mu, \sigma^2) &= \frac{1}{2\pi\sigma^2} \exp\{-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\} \\ &= h(x)g(\boldsymbol{\theta})exp\{\boldsymbol{\theta}^\top \boldsymbol{u}(\boldsymbol{x})\} \end{aligned}$$

Conjugate priors

Exponential family distributions

- A large class of probability distributions are part of the exponential family (all in this course) and can be written as:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x})g(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^\top \boldsymbol{u}(\boldsymbol{x})\}$$

- For example for the Gaussian:

$$\begin{aligned} p(x \mid \mu, \sigma^2) &= \frac{1}{2\pi\sigma^2} \exp\{-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\} \\ &= h(x)g(\boldsymbol{\theta}) exp\{\boldsymbol{\theta}^\top \boldsymbol{u}(\boldsymbol{x})\} \end{aligned}$$

Conjugate priors

Exponential family distributions

- A large class of probability distributions are part of the exponential family (all in this course) and can be written as:

$$p(\boldsymbol{x} \mid \boldsymbol{\theta}) = h(\boldsymbol{x})g(\boldsymbol{\theta}) \exp\{\boldsymbol{\theta}^\top \boldsymbol{u}(\boldsymbol{x})\}$$

- For example for the Gaussian:

$$\begin{aligned} p(x \mid \mu, \sigma^2) &= \frac{1}{2\pi\sigma^2} \exp\{-\frac{1}{2\sigma^2}(x^2 - 2x\mu + \mu^2)\} \\ &= h(x)g(\boldsymbol{\theta})exp\{\boldsymbol{\theta}^\top \boldsymbol{u}(\boldsymbol{x})\} \end{aligned}$$

with $\boldsymbol{\theta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix}$, $h(x) = \dfrac{1}{\sqrt{2\pi}}$

$\boldsymbol{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix}$, $g(\boldsymbol{\theta}) = (-2\theta_2)^{1/2} \exp\left(\dfrac{\theta_1^2}{4\theta_2}\right)$

Conjugacy and exponential family distributions

- For all members of the exponential family it is possible to construct a conjugate prior.
  - Intuition: The exponential form ensures that we can construct a prior that keeps its functional form.

- Conjugate priors for the Gaussian $\mathcal{N}\left(x \mid \mu, \sigma^2\right)$
  - $p(\mu) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right)$
  - $p(\frac{1}{\sigma^2}) = \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$.
  - $p(\mu, \frac{1}{\sigma^2}) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right) \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$

Conjugacy and exponential family distributions

- For all members of the exponential family it is possible to construct a conjugate prior.
  - Intuition: The exponential form ensures that we can construct a prior that keeps its functional form.

- Conjugate priors for the Gaussian $\mathcal{N}\left(x \mid \mu, \sigma^2\right)$
  - $p(\mu) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right)$
  - $p(\frac{1}{\sigma^2}) = \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$.
  - $p(\mu, \frac{1}{\sigma^2}) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right) \cdot \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$

Conjugacy and exponential family distributions

- For all members of the exponential family it is possible to construct a conjugate prior.
  - Intuition: The exponential form ensures that we can construct a prior that keeps its functional form.

- Conjugate priors for the Gaussian $\mathcal{N}\left(x \mid \mu, \sigma^2\right)$
  - $p(\mu) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right)$
  - $p(\frac{1}{\sigma^2}) = \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$.
  - $p(\mu, \frac{1}{\sigma^2}) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right) \cdot \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$

Conjugacy and exponential family distributions

- For all members of the exponential family it is possible to construct a conjugate prior.
  - Intuition: The exponential form ensures that we can construct a prior that keeps its functional form.

- Conjugate priors for the Gaussian $\mathcal{N}\left(x \mid \mu, \sigma^2\right)$
  - $p(\mu) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right)$
  - $p(\frac{1}{\sigma^2}) = \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$.
  - $p(\mu, \frac{1}{\sigma^2}) = \mathcal{N}\left(\mu \mid m_0,\ s_0^2\right) \cdot \mathcal{G}\left(\frac{1}{\sigma^2} \mid a_0, b_0\right)$

Gamma distribution

$$\mathcal{G}(x \mid a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}$$

Bayesian Inference for the Gaussian

Sequential learning

- Bayes rule naturally leads itself to sequential learning
- Assume one by one multiple datasets become available: $\mathcal{D}_1, \ldots, \mathcal{D}_S$

$$p_1(\boldsymbol{\theta}) \propto p(\mathcal{D}_1 \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$
$$p_2(\boldsymbol{\theta}) \propto p(\mathcal{D}_2 \mid \boldsymbol{\theta}) p_1(\boldsymbol{\theta})$$
$$\ldots$$

- Note: Assuming the datasets are independent, sequential updates and a single learning step yield the same answer.

Bayesian Inference for the Gaussian

Sequential learning

- Bayes rule naturally leads itself to sequential learning
- Assume one by one multiple datasets become available: $\mathcal{D}_1, \ldots, \mathcal{D}_S$

$$p_1(\boldsymbol{\theta}) \propto p(\mathcal{D}_1 \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})$$
$$p_2(\boldsymbol{\theta}) \propto p(\mathcal{D}_2 \mid \boldsymbol{\theta}) p_1(\boldsymbol{\theta})$$
$$\cdots$$

- Note: Assuming the datasets are independent, sequential updates and a single learning step yield the same answer.

Bayesian Inference for the Gaussian

Sequential learning

- Bayes rule naturally leads itself to sequential learning
- Assume one by one multiple datasets become available: $\mathcal{D}_1, \ldots, \mathcal{D}_S$

$$
\begin{aligned}
p_1(\boldsymbol{\theta}) &\propto p(\mathcal{D}_1 \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\
p_2(\boldsymbol{\theta}) &\propto p(\mathcal{D}_2 \,|\, \boldsymbol{\theta}) p_1(\boldsymbol{\theta}) \\
&\cdots
\end{aligned}
$$

- Note: Assuming the datasets are independent, sequential updates and a single learning step yield the same answer.

Outline

Regression

Noise model and likelihood

- Given a dataset $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{S}$, where $\boldsymbol{x}_n = \{x_{n,1}, \ldots, x_{n,S}\}$ is S dimensional, fit parameters $\boldsymbol{\theta}$ of a regressor f with added Gaussian noise:

$$y_n = f(\boldsymbol{x}_n; \boldsymbol{\theta}) + \epsilon_n \quad \text{where} \quad p(\epsilon \,|\, \sigma^2) = \mathcal{N}\left(\epsilon \mid 0, \sigma^2\right).$$

- Equivalent likelihood formulation:

$$p(\boldsymbol{y} \,|\, \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid f(\boldsymbol{x}_n; \boldsymbol{\theta}), \sigma^2\right)$$

- Choose f to be linear:

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta} + c, \sigma^2 \right)$$

- Consider bias free case, $c = 0$, otherwise include an additional column of ones in each $\boldsymbol{x}_n$.

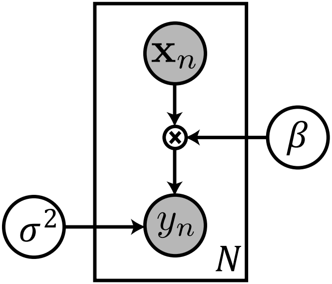
Regression

Choosing a regressor

- Choose f to be linear:

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = \prod_{n=1}^{N} \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta} + c, \sigma^2 \right)$$

- Consider bias free case, $c = 0$, otherwise include an additional column of ones in each $\boldsymbol{x}_n$.



Equivalent graphical model

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}\sigma^2) = \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right)$$

$$= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2}_{\text{Sum of squares}}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

Maximum likelihood

- Taking the logarithm, we obtain

$$
\begin{aligned}
\ln p(\boldsymbol{y} \mid \boldsymbol{\theta}\sigma^2) &= \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right) \\
&= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2}_{\text{Sum of squares}}
\end{aligned}
$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression

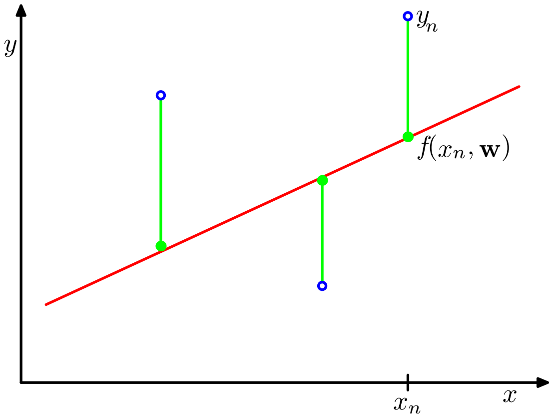
Maximum likelihood

- Taking the logarithm, we obtain

$$\begin{aligned}\ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}\sigma^2) &= \sum_{n=1}^{N} \ln \mathcal{N}\left(y_n \mid \boldsymbol{x}_n \cdot \boldsymbol{\beta}, \sigma^2 \right) \\ &= -\frac{N}{2} \ln 2\pi\sigma^2 - \frac{1}{2\sigma^2} \underbrace{\sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2}_{\text{Sum of squares}}\end{aligned}$$

- The likelihood is maximized when the squared error is minimized.
- Least squares and maximum likelihood are equivalent.

Linear Regression and Least Squares



(C.M. Bishop, Pattern Recognition and Machine Learning)

$$E(\boldsymbol{\beta}) = \frac{1}{2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2$$

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_i} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\theta})^2 \right]$$

- Set gradient w.r.t. β to zero

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_i} \ln p(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\theta})^2 \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero

Linear Regression and Least Squares

- Derivative w.r.t a single weight entry β_i

$$\frac{d}{\mathrm{d}\beta_i} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\theta}, \sigma^2) = \frac{d}{\mathrm{d}\beta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\theta})^2 \right]$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \boldsymbol{x}_n^\top = 0$$

$$\implies \boldsymbol{\beta}_{\mathrm{M}} = ?$$

Linear Regression and Least Squares

- Derivative w.r.t. a single weight entry β_i

$$\frac{\partial}{\partial \beta_i} \ln p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = \frac{\partial}{\partial \beta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$
$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) x_i$$

- Set gradient w.r.t. β to zero

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \boldsymbol{x}_n^\top = 0$$
$$\implies \beta_{\mathrm{M}} = \underbrace{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{\text{Pseudo inverse}} \boldsymbol{y}$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \dots & x1, S \\ \dots & \dots & \dots \\ x_{N,1} & \dots & x_{N,S} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t. a single weight entry β_i

$$\frac{\partial}{\partial \beta_i} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{\partial}{\partial \beta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$
$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) x_i$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \boldsymbol{x}_n^\top = 0$$
$$\implies \boldsymbol{\beta}_{\mathrm{M}} = \underbrace{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{\text{Pseudo inverse}} \boldsymbol{y}$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \dots & x1,S \\ \dots & \dots & \dots \\ x_{N,1} & \dots & x_{N,S} \end{bmatrix}$

Linear Regression and Least Squares

- Derivative w.r.t. a single weight entry β_i

$$\frac{\partial}{\partial \beta_i} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{\partial}{\partial \beta_i} \left[-\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta})^2 \right]$$
$$= \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) x_i$$

- Set gradient w.r.t. $\boldsymbol{\beta}$ to zero

$$\nabla_{\boldsymbol{\beta}} \ln p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} (y_n - \boldsymbol{x}_n \cdot \boldsymbol{\beta}) \boldsymbol{x}_n^\top = 0$$
$$\implies \boldsymbol{\beta}_{\mathrm{M}} = \underbrace{(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top}_{\text{Pseudo inverse}} \boldsymbol{y}$$

- Here, the matrix $\boldsymbol{X}$ is defined as $\boldsymbol{X} = \begin{bmatrix} x_{1,1} & \dots & x1, S \\ \dots & \dots & \dots \\ x_{N,1} & \dots & x_{N,S} \end{bmatrix}$

Outline

Conclusions
Summary - week 1

- Probability theory: the language of uncertainty.
- Key rules of probability: sum rule, product rule.
- Bayes rules formes the fundamentals of learning. (posterior $\propto$ likelihood $\cdot$ prior).
- The entropy quantifies uncertainty.
- Parameter learning using maximum likelihood.
- Bayesian inference for the Gaussian.
- Linear regression
  - Parameter estimation
  -

Conclusions

Summary - week 1

- Probability theory: the language of uncertainty.
- Key rules of probability: sum rule, product rule.
- Bayes rules formes the fundamentals of learning. (posterior $\propto$ likelihood $\cdot$ prior).
- The entropy quantifies uncertainty.
- Parameter learning using maximum likelihood.
- Bayesian inference for the Gaussian.
- Linear regression
  - Parameter estimation
  -

Outlook - week 2

- Revisit the (multivariate) normal distribution, showing some useful properties.
- Statistical testing
- Genome-wide association studies using linear regression
- Bayesian linear regression and shrinkage

Conclusions

Summary - week 1

- Probability theory: the language of uncertainty.
- Key rules of probability: sum rule, product rule.
- Bayes rules formes the fundamentals of learning. (posterior $\propto$ likelihood $\cdot$ prior).
- The entropy quantifies uncertainty.
- Parameter learning using maximum likelihood.
- Bayesian inference for the Gaussian.
- Linear regression
  - Parameter estimation
  -

Outlook - week 2

- Revisit the (multivariate) normal distribution, showing some useful properties.
- Statistical testing
- Genome-wide association studies using linear regression
- Bayesian linear regression and shrinkage

Acknowledgements

- Oliver Stegle (builds on joint course material)