Introduction to Artificial Intelligence

Ulf Brefeld & Eraldo R. Fernandes
build: April 12, 2022

Machine Learning Group
Leuphana University of Lüneburg

Course General Information

Course General Information

- Lecturers
  - Ulf Brefeld (Lecture)
  - Eraldo R. Fernandes (Tutorial)
- Lecture: Tue 14:15–15:45 @ HS 5
- Tutorial: Thu 9:15, 10:15, 11:15 @ C11.320
- Consultation hours on request (just send an email)
  - brefeld@leuphana.de
  - eraldo.fernandes@leuphana.de

Ulf Brefeld

Eraldo R. Fernandes

Grading and Tutorial

- Written exam (see myStudy)
- Tutorials
  - Regular exercise sheets ($\sim$ weekly)
  - Practical questions
  - Discussions
- Each tutorial
  - You mark all tasks you worked on (solved or tried to solve)
  - We discuss solutions together
- You need 50% of the marks to pass the tutorial

Recap Overview

- Calculus
- Linear Algebra
- Probability Theory
- Supervised Machine Learning

Calculus Recap

Sum and Product

- Sequence of n numbers: $(a_1, a_2, \ldots, a_n) = (a_i)_{i=1}^{n}$

Sum and Product

- Sequence of n numbers: $(a_1, a_2, \dots, a_n) = (a_i)_{i=1}^n$
- Sum: $a_1 + a_2 + \cdots + a_n$
  - $\sum_{i=1}^n a_i$
  - $\sum_{i \in I} a_i$ for $I = \{1, 2, \dots, n\}$
  - $\sum_{a \in A} a$, for $A = \{a_1, a_2, \dots, a_n\}$
- For $I = \emptyset$ or $A = \emptyset$: the sum is equal to zero

Sum and Product

- Sequence of n numbers: $(a_1, a_2, \ldots, a_n) = (a_i)_{i=1}^n$
- Sum: $a_1 + a_2 + \cdots + a_n$
  - $\sum_{i=1}^n a_i$
  - $\sum_{i \in I} a_i$ for $I = \{1, 2, \ldots, n\}$
  - $\sum_{a \in A} a$, for $A = \{a_1, a_2, \ldots, a_n\}$
- For $I = \emptyset$ or $A = \emptyset$: the sum is equal to zero
- Product: $a_1 \cdot a_2 \cdot \ldots \cdot a_n$
  - $\prod_{i=1}^n a_i = a_1 \cdot a_2 \cdot a_3 \cdot \ldots \cdot a_n$
- For $I = \emptyset$ or $A = \emptyset$, the product is equal to one

Binomial Coefficient

- *Factorial* of $n \in \mathbb{N}$

$$n! = \prod_{k=1}^{n} k = 1 \cdot 2 \dots k$$

Binomial Coefficient

- *Factorial* of $n \in \mathbb{N}$

$$n! = \prod_{k=1}^{n} k = 1 \cdot 2 \ldots k$$

  - $0! = 1$
  - Example: $6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720$

Binomial Coefficient

- *Factorial* of $n \in \mathbb{N}$
$$n! = \prod_{k=1}^{n} k = 1 \cdot 2 \ldots k$$
  - $0! = 1$
  - Example: $6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720$
- The *binomial coefficient* for $n, k \in \mathbb{Z}^+$
$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

Binomial Coefficient

- *Factorial* of $n \in \mathbb{N}$

$$n! = \prod_{k=1}^{n} k = 1 \cdot 2 \ldots k$$

  - $0! = 1$
  - Example: $6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720$

- The *binomial coefficient* for $n, k \in \mathbb{Z}^+$

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

  - Example:

$$\binom{6}{2} = \frac{6!}{2! \cdot (6-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3 \cdot 4} = \frac{5 \cdot 6}{1 \cdot 2} = \frac{30}{2} = 15$$

Binomial Coefficient

- *Factorial* of $n \in \mathbb{N}$

$$n! = \prod_{k=1}^{n} k = 1 \cdot 2 \ldots k$$

  - $0! = 1$
  - Example: $6! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 = 720$
- The *binomial coefficient* for $n, k \in \mathbb{Z}^+$

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!}$$

  - Example:

$$\binom{6}{2} = \frac{6!}{2! \cdot (6-2)!} = \frac{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6}{1 \cdot 2 \cdot 1 \cdot 2 \cdot 3 \cdot 4} = \frac{5 \cdot 6}{1 \cdot 2} = \frac{30}{2} = 15$$

  - $\binom{n}{k}$ = number of combinations with k elements from a set with n elements

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$
- Sum rule: $(f + g)'(x) = f'(x) + g'(x)$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$
- Sum rule: $(f + g)'(x) = f'(x) + g'(x)$
- Product rule: $(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$
- Sum rule: $(f + g)'(x) = f'(x) + g'(x)$
- Product rule: $(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$
- Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g^2(x)}$ für $g(x) \neq 0$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$
- Sum rule: $(f + g)'(x) = f'(x) + g'(x)$
- Product rule: $(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$
- Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g^2(x)}$ für $g(x) \neq 0$
- Chain rule: $(g \circ f)' = g(f(x))' = g'(f(x)) \cdot f'(x)$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$
- Sum rule: $(f + g)'(x) = f'(x) + g'(x)$
- Product rule: $(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$
- Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g^2(x)}$ für $g(x) \neq 0$
- Chain rule: $(g \circ f)' = g(f(x))' = g'(f(x)) \cdot f'(x)$
- Example
    - $f(x) = a \cdot x^2 + b \cdot x + c$

Derivative Rules

- Let f, g be differentiable functions of $x \in \mathbb{R}$ and $\lambda \in \mathbb{R}$
- $f(x)' = f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} f(x)$
- Constant multiplication: $(\lambda \cdot f(x))' = \lambda \cdot f'(x)$
- Sum rule: $(f+g)'(x) = f'(x) + g'(x)$
- Product rule: $(f \cdot g)'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$
- Quotient rule: $\left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x) \cdot g(x) - f(x) \cdot g'(x)}{g^2(x)}$ für $g(x) \neq 0$
- Chain rule: $(g \circ f)' = g(f(x))' = g'(f(x)) \cdot f'(x)$
- Example
  - $f(x) = a \cdot x^2 + b \cdot x + c$

$$\begin{aligned} f'(x) &= \left(a \cdot x^2 + b \cdot x + c\right)' \\ &= \left(a \cdot x^2\right)' + (b \cdot x)' + c' \\ &= a \cdot \left(x^2\right)' + b \cdot x' + c' \\ &= 2ax + b \end{aligned}$$

Integration Rules

- Linearity

$$\int (f+g)(x)\mathrm{d}x = \int f(x)\mathrm{d}x + \int g(x)\mathrm{d}x$$

$$\int (\lambda f)(x)\mathrm{d}x = \lambda \cdot \int f(x)\mathrm{d}x$$

Integration Rules

- Linearity

$$\int (f+g)(x)\mathrm{d}x = \int f(x)\mathrm{d}x + \int g(x)\mathrm{d}x$$

$$\int (\lambda f)(x)\mathrm{d}x = \lambda \cdot \int f(x)\mathrm{d}x$$

- Definite integral (for $a < b < c$)

$$\int_a^c f(x)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_b^c f(x)\mathrm{d}x$$

Integration Rules

- Linearity

$$\int (f+g)(x)\mathrm{d}x = \int f(x)\mathrm{d}x + \int g(x)\mathrm{d}x$$

$$\int (\lambda f)(x)\mathrm{d}x = \lambda \cdot \int f(x)\mathrm{d}x$$

- Definite integral (for $a < b < c$)

$$\int_a^c f(x)\mathrm{d}x = \int_a^b f(x)\mathrm{d}x + \int_b^c f(x)\mathrm{d}x$$

  - Primitive (antiderivative) function: $F' = f$

$$\int_a^b f(x)\mathrm{d}x = [F(x)]_a^b = F(b) - F(a)$$

Integration Rules – Example

- $f(x) = a \cdot x^2 + b \cdot x + c$
- $u, v \in \mathbb{R}$

Integration Rules – Example

- $f(x) = a \cdot x^2 + b \cdot x + c$
- $u, v \in \mathbb{R}$

$$
\begin{aligned}
\int_u^v f(x)\mathrm{d}x &= \int_u^v \left(a \cdot x^2 + b \cdot x + c\right) \mathrm{d}x \\
&= \int_u^v a \cdot x^2 \mathrm{d}x + \int_u^v b \cdot x \mathrm{d}x + \int_u^v c \mathrm{d}x \\
&= a \cdot \int_u^v x^2 \mathrm{d}x + b \cdot \int_u^v x \mathrm{d}x + c \cdot \int_u^v 1 \mathrm{d}x \\
&= a \cdot \left[\frac{1}{3}x^3\right]_u^v + b \cdot \left[\frac{1}{2}x^2\right]_u^v + c \cdot [x]_u^v \\
&= \frac{a}{3}\left(v^3 - u^3\right) + \frac{b}{2}\left(v^2 - u^2\right) + c\left(v - u\right)
\end{aligned}
$$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$
- $a^{x+y} = a^x \cdot a^y$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$
- $a^{x+y} = a^x \cdot a^y$
- $a^{x-y} = \frac{a^x}{a^y}$ für $a^y \neq 0$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$
- $a^{x+y} = a^x \cdot a^y$
- $a^{x-y} = \frac{a^x}{a^y}$ für $a^y \neq 0$
- $a^{x \cdot y} = (a^x)^y$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$
- $a^{x+y} = a^x \cdot a^y$
- $a^{x-y} = \frac{a^x}{a^y}$ für $a^y \neq 0$
- $a^{x \cdot y} = (a^x)^y$
- $a^x \cdot b^x = (a \cdot b)^x$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$
- $a^{x+y} = a^x \cdot a^y$
- $a^{x-y} = \frac{a^x}{a^y}$ für $a^y \neq 0$
- $a^{x \cdot y} = (a^x)^y$
- $a^x \cdot b^x = (a \cdot b)^x$
- For $a = e$, we use $\exp(x) = e^x$

Exponential Function

$a, b, x, y \in \mathbb{R}$

- $a^0 = 1$
- $a^1 = a$
- $a^{x+y} = a^x \cdot a^y$
- $a^{x-y} = \frac{a^x}{a^y}$ für $a^y \neq 0$
- $a^{x \cdot y} = (a^x)^y$
- $a^x \cdot b^x = (a \cdot b)^x$
- For $a = e$, we use $\exp(x) = e^x$
- $\exp(x)' = \exp(x)$

Logarithm Function

- $\log(x \cdot y) = \log x + \log y$

Logarithm Function

- $\log(x \cdot y) = \log x + \log y$
- $\log\left(\prod_{k=0}^{n} x_k\right) = \sum_{k=0}^{n} \log x_k$

Logarithm Function

- $\log(x \cdot y) = \log x + \log y$
- $\log\left(\prod_{k=0}^{n} x_k\right) = \sum_{k=0}^{n} \log x_k$
- $\log\left(\frac{x}{y}\right) = \log x - \log y$

Logarithm Function

- $\log(x \cdot y) = \log x + \log y$
- $\log\left(\prod_{k=0}^{n} x_k\right) = \sum_{k=0}^{n} \log x_k$
- $\log\left(\frac{x}{y}\right) = \log x - \log y$
- $\log(x^r) = r \cdot \log x$

Logarithm Function

- $\log(x \cdot y) = \log x + \log y$
- $\log\left(\prod_{k=0}^{n} x_k\right) = \sum_{k=0}^{n} \log x_k$
- $\log\left(\frac{x}{y}\right) = \log x - \log y$
- $\log(x^r) = r \cdot \log x$
- $\log'(x) = \frac{1}{x}$

Logarithm Function

- $\log(x \cdot y) = \log x + \log y$
- $\log\left(\prod_{k=0}^{n} x_k\right) = \sum_{k=0}^{n} \log x_k$
- $\log\left(\frac{x}{y}\right) = \log x - \log y$
- $\log(x^r) = r \cdot \log x$
- $\log'(x) = \frac{1}{x}$
- Natural logarithm: $\ln(x) = \log_a(x)$ when $a = e$

Linear Algebra Recap

Vectors

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \in \mathbb{R}^d$$

Vectors

- Multiplication by a scalar $c \in \mathbb{R}$

$$c\mathbf{x} = \begin{pmatrix} cx_1 \\ cx_2 \\ \vdots \\ cx_d \end{pmatrix}$$

- Addition of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x} + \mathbf{y} = \begin{pmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_d + y_d \end{pmatrix}$$

- Transpose

$$\mathbf{x}^\top = \begin{pmatrix} x_1 & x_2 & \dots & x_d \end{pmatrix}$$

Vectors

- Inner product of two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\mathbf{x}^\top \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + \ldots + x_d y_d = \sum_{i=1}^{d} x_i y_i \in \mathbb{R}$$

- Outer product of two vectors $\mathbf{x} \in \mathbb{R}^{d_1}, \mathbf{y} \in \mathbb{R}^{d_2}$

$$\mathbf{x}\mathbf{y}^\top = \begin{pmatrix} x_1 y_1 & \ldots & x_1 y_{d_2} \\ \vdots & \ddots & \vdots \\ x_{d_1} y_1 & \ldots & x_{d_1} y_{d_2} \end{pmatrix} \in \mathbb{R}^{d_1 \times d_2}$$

Vectors

- Length of a vector $\mathbf{x} \in \mathbb{R}^d$

$$\mathbf{x} = \|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^\top \mathbf{x}} = \sqrt{x_1^2 + \ldots + x_d^2} \in \mathbb{R}$$

- Angle between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$

$$\cos(\theta) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \cdot \|\mathbf{y}\|_2}$$

Matrices

A matrix can be seen as a collection of vectors

$$\mathbf{A} = \begin{pmatrix} a_{11} & \dots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nm} \end{pmatrix} = \begin{pmatrix} | & | & | \\ \mathbf{a}_1 & \dots & \mathbf{a}_m \\ | & | & | \end{pmatrix} \in \mathbb{R}^{n \times m}$$

- Transpose

$$\mathbf{A}^\top = \begin{pmatrix} a_{11} & \dots & a_{n1} \\ \vdots & \ddots & \vdots \\ a_{1m} & \dots & a_{nm} \end{pmatrix} \in \mathbb{R}^{m \times n}$$

Matrices

- Multiplication by a scalar $c \in \mathbb{R}$

$$c\mathbf{A} = \begin{pmatrix} ca_{11} & \dots & ca_{1m} \\ \vdots & \ddots & \vdots \\ ca_{n1} & \dots & ca_{nm} \end{pmatrix}$$

- Addition of two matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times m}$

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} a_{11} + b_{11} & \dots & a_{1m} + b_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} + b_{n1} & \dots & a_{nm} + b_{nm} \end{pmatrix}$$

Matrices

- Multiplication of a vector $\mathbf{x} \in \mathbb{R}^m$ by a matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$

$$\mathbf{A}\mathbf{x} = \mathbf{b} \in \mathbb{R}^n$$

$$b_i = \sum_{j=1}^{m} a_{ij} x_j$$

- Multiplication of two matrices $\mathbf{A} \in \mathbb{R}^{n_1 \times m}$ and $\mathbf{B} \in \mathbb{R}^{m \times n_2}$

$$\mathbf{A}\mathbf{B} = \mathbf{C} \in \mathbb{R}^{n_1 \times n_2}$$

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}$$

Matrices

- Inverse of a matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

Whenever you can avoid calculating the inverse of a matrix AVOID IT! E.g. solving the linear equation system $\mathbf{A}\mathbf{x} = \mathbf{b}$ for $\mathbf{x}$ is much better than working with the inverse: $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$

Norm

- Norm: function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_0^+$
- Properties ($x, y \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$)
  - $\|x\| = 0 \implies x = 0$
  - $\|\lambda x\| = |\lambda| \cdot \|x\|$
  - $\|x + y\| \leq \|x\| + \|y\|$
- p-Norm ($p \geq 1$)

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

Norm

- Norm: function $\|\cdot\| : \mathbb{R}^d \to \mathbb{R}_0^+$
- Properties ($x, y \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$)
  - $\|x\| = 0 \implies x = 0$
  - $\|\lambda x\| = |\lambda| \cdot \|x\|$
  - $\|x + y\| \leq \|x\| + \|y\|$
- p-Norm ($p \geq 1$)

$$\|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{\frac{1}{p}}$$

- 1-Norm: $\|x\|_1 = \sum_{i=1}^{d} |x_i|$.
- Euclidean norm (2-norm): $\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$.
- Maximum norm ($p \to \infty$): $\|x\|_\infty = \max_i |x_i|$.

Probability Theory Recap

Probability Theory

- Why probabilities?
- Of course, in case of random events, stochasticity, ...

Probability Theory

- Why probabilities?
- Of course, in case of random events, stochasticity, ...
- Additionally
  - Lack of knowledge
  - Hidden (latent) variables
  - Expressing uncertainty
  - Expressing information

Probability Theory

- Why probabilities?
- Of course, in case of random events, stochasticity, ...
- Additionally
  - Lack of knowledge
  - Hidden (latent) variables
  - Expressing uncertainty
  - Expressing information
- Generic tool to express uncertainty, information, and coupling

(Discrete) Random Variables

- Intuitively: probability of *random variable* X taking on value x
- *Example*
  - A dice roll (X) can result in $\{1, \ldots, 6\}$
  - What is the probability of $X = x$ for $x \in \{1, \ldots, 6\}$

(Discrete) Random Variables

- Intuitively: probability of *random variable* X taking on value x
- *Example*
    - A dice roll (X) can result in $\{1, \ldots, 6\}$
    - What is the probability of $X = x$ for $x \in \{1, \ldots, 6\}$
- A bit more formally
    - Domain of X: $\text{dom}(X)$ (or *sample space* Ω)
        - Set of possible values of a random variable
        - Mutually exclusive: only one will happen
        - Collectively exhaustive: at least one will happen

(Discrete) Random Variables

- Intuitively: probability of *random variable* X taking on value x
- *Example*
  - A dice roll (X) can result in $\{1, \ldots, 6\}$
  - What is the probability of $X = x$ for $x \in \{1, \ldots, 6\}$
- A bit more formally
  - Domain of X: $\text{dom}(X)$ (or *sample space* Ω)
    - Set of possible values of a random variable
    - Mutually exclusive: only one will happen
    - Collectively exhaustive: at least one will happen
  - Space of events $\mathcal{F} = \{E : E \subset \Omega\}$ and $A, B, \emptyset, \Omega \in \mathcal{F}$
    - For event $E \in \mathcal{F}$, $P(E) = \sum_{x \in E} P(X = x)$
    - $P(A) \in [0, 1]$
    - $P(\Omega) = 1$ (sure event)
    - $P(\emptyset) = 0$ (impossible event)
    - If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$

(Discrete) Random Variables

- Notation
  - Random variable X (capital letter)
  - Value $x \in \text{dom}(X)$ is taken by X (lower case letter)
  - $P(X = x) \in \mathbb{R}$: $x \in \text{dom}(X) \to$ probability in $[0, 1]$
  - For event $E \subset \Omega$
    - $P(E)$: probability of X taking a value $x \in E$
    - $P(E) = \sum_{x \in E} P(X = x)$

Joint Probability Distributions

- Two random variables: X, Y
- Joint probability distribution: $P(X = x, Y = y)$
  - Probability that $X = x$ and $Y = y$

Joint Probability Distributions

- Two random variables: X, Y
- Joint probability distribution: $P(X = x, Y = y)$
  - Probability that $X = x$ and $Y = y$
- In logic: $X = x \wedge Y = y$
  - Not so in joint probability distributions

Joint Probability Distributions

- Two random variables: X, Y
- Joint probability distribution: $P(X = x, Y = y)$
  - Probability that $X = x$ and $Y = y$
- In logic: $X = x \wedge Y = y$
  - Not so in joint probability distributions

Example Two binary RV: Cavity (X) and Toothache (Y)

$P(X, Y)$	$Y = T$	$Y = F$
$X = T$	0.04	0.06
$X = F$	0.01	0.89

Joint Probability Distributions

- Two random variables: X, Y
- Joint probability distribution: $P(X = x, Y = y)$
  - Probability that $X = x$ and $Y = y$
- In logic: $X = x \wedge Y = y$
  - Not so in joint probability distributions

Example Two binary RV: Cavity (X) and Toothache (Y)

$P(X, Y)$	$Y = T$	$Y = F$
$X = T$	0.04	0.06
$X = F$	0.01	0.89

We write: $P(X = F, Y = T) = 0.01$

Joint Probability Distributions – Definitions

Marginal probability of X given $P(X, Y)$

$$P(X) = \sum_{Y} P(X, Y)$$

$P(X, Y)$	$Y = T$	$Y = F$	$P(X)$
$X = T$	0.15	0.20	0.35
$X = F$	0.05	0.60	0.65
$P(Y)$	0.20	0.80	

Joint Probability Distributions – Definitions

Marginal probability of X given $P(X, Y)$

$$P(X) = \sum_Y P(X, Y)$$

Conditional probability of X given $P(Y)$ and $P(X, Y)$

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

$P(X, Y)$	$Y = T$	$Y = F$	$P(X)$
$X = T$	0.15	0.20	0.35
$X = F$	0.05	0.60	0.65
$P(Y)$	0.20	0.80	

	$P(X\|Y = T)$	$P(X\|Y = F)$
$X = T$	0.75	0.25
$X = F$	0.25	0.75

Joint Probability Distributions – Definitions

Marginal probability of X given $P(X, Y)$

$$P(X) = \sum_Y P(X, Y)$$

$P(X, Y)$	$Y = T$	$Y = F$	$P(X)$
$X = T$	0.15	0.20	0.35
$X = F$	0.05	0.60	0.65
$P(Y)$	0.20	0.80	

Conditional probability of X given $P(Y)$ and $P(X, Y)$

$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

	$P(X\|Y = T)$	$P(X\|Y = F)$
$X = T$	0.75	0.25
$X = F$	0.25	0.75

Extends for more than two variables (vector of variables)

$$X = (X_1, \dots, X_n)$$
$$P(X_1, \dots, X_{n-1}, X_n)$$
$$P(X_n | X_1, \dots, X_{n-1})$$

Joint Probability Distributions – Properties

- Product rule

$$P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

Joint Probability Distributions – Properties

- Product rule

$$P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

- Chain rule

$$P(X_1, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1})\ P(X_1, \ldots, X_{n-1})$$

Joint Probability Distributions – Properties

- Product rule

$$P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

- Chain rule

$$P(X_1, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1})\ P(X_1, \ldots, X_{n-1})$$

$$= P(X_n|X_1, \ldots, X_{n-1})\ P(X_{n-1}|X_1, \ldots, X_{n-2})\ P(X_1, \ldots, X_{n-2})$$

Joint Probability Distributions – Properties

- Product rule

$$P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

- Chain rule

$$P(X_1, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1})\ P(X_1, \ldots, X_{n-1})$$

$$= P(X_n|X_1, \ldots, X_{n-1})\ P(X_{n-1}|X_1, \ldots, X_{n-2})\ P(X_1, \ldots, X_{n-2})$$

$$= P(X_n|X_1, \ldots, X_{n-1})\ P(X_{n-1}|X_1, \ldots, X_{n-2})\ \cdots\ P(X_1)$$

Joint Probability Distributions – Properties

- Product rule

$$P(X,Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

- Chain rule

$$P(X_1,\ldots,X_n) = P(X_n|X_1,\ldots,X_{n-1})\ P(X_1,\ldots,X_{n-1})$$

$$= P(X_n|X_1,\ldots,X_{n-1})\ P(X_{n-1}|X_1,\ldots,X_{n-2})\ P(X_1,\ldots,X_{n-2})$$

$$= P(X_n|X_1,\ldots,X_{n-1})\ P(X_{n-1}|X_1,\ldots,X_{n-2})\ \cdots\ P(X_1)$$

$$= \prod_{i=1}^{n} P(X_i|X_1,\ldots,X_{i-1}) = \prod_{i=1}^{n} P(X_i|X_{j<i})$$

Joint Probability Distributions – Properties

- Product rule

$$P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

- Chain rule

$$P(X_1, \ldots, X_n) = P(X_n|X_1, \ldots, X_{n-1})\ P(X_1, \ldots, X_{n-1})$$

$$= P(X_n|X_1, \ldots, X_{n-1})\ P(X_{n-1}|X_1, \ldots, X_{n-2})\ P(X_1, \ldots, X_{n-2})$$

$$= P(X_n|X_1, \ldots, X_{n-1})\ P(X_{n-1}|X_1, \ldots, X_{n-2})\ \cdots\ P(X_1)$$

$$= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) = \prod_{i=1}^{n} P(X_i|X_{j<i})$$

- Bayes rule

$$P(Y|X) = \frac{P(X|Y)}{P(X)} P(Y)$$

Joint Probability Distributions

- Most of what we will need involves
    - Joint Probability Distributions

Joint Probability Distributions

- Most of what we will need involves
  - Joint Probability Distributions
- Graphical models
  - Descriptions of joint probability distributions

Joint Probability Distributions

- Most of what we will need involves
  - Joint Probability Distributions
- Graphical models
  - Descriptions of joint probability distributions
- Correlations, interdependence, and coupling
  - Expressed in terms of joint probability distributions

Bayes Rule

- Thomas Bayes (1702–1761)
- Trivial implication of marginal and conditional probability
- Important: interpretation and use

$$P(Y|X) = \frac{P(X|Y)}{P(X)} \, P(Y)$$

Bayes Rule

- Thomas Bayes (1702–1761)
- Trivial implication of marginal and conditional probability
- Important: interpretation and use

$$P(Y|X) = \frac{P(X|Y)}{P(X)} \, P(Y), \quad \text{posterior} = \frac{\text{likelihood}}{\text{evidence}} \, \text{prior}$$

Bayes Rule

- Thomas Bayes (1702–1761)
- Trivial implication of marginal and conditional probability
- Important: interpretation and use

$$P(Y|X) = \frac{P(X|Y)}{P(X)}\, P(Y), \quad \text{posterior} = \frac{\text{likelihood}}{\text{evidence}}\, \text{prior}$$

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})}{P(\text{effect})}\, P(\text{cause})$$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
    - 95% sensitivity and specificity
    - 5% chance of being wrong (both sides)
        - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = ?$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos},\text{ab}) + P(\text{pos},!\text{ab})}$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos,ab}) + P(\text{pos,!ab})}$
- Product rule: $P(\text{pos}, A) = P(\text{pos}|A)\ P(A)$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos},\text{ab}) + P(\text{pos},!\text{ab})}$
- Product rule: $P(\text{pos}, A) = P(\text{pos}|A)\ P(A)$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos}|\text{ab})P(\text{ab}) + P(\text{pos}|!\text{ab})P(!\text{ab})}$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos},\text{ab}) + P(\text{pos},!\text{ab})}$
- Product rule: $P(\text{pos}, A) = P(\text{pos}|A)\ P(A)$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos}|\text{ab})P(\text{ab}) + P(\text{pos}|!\text{ab})P(!\text{ab})}$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos,ab}) + P(\text{pos,!ab})}$
- Product rule: $P(\text{pos}, A) = P(\text{pos}|A)\ P(A)$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos}|\text{ab})P(\text{ab}) + P(\text{pos}|!\text{ab})P(!\text{ab})}$
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{0.95 \cdot 0.05 + 0.05 \cdot 0.95}$

Bayesian Inference – Example (by @xamat)

- Bad COVID antibody test kit
  - 95% sensitivity and specificity
  - 5% chance of being wrong (both sides)
    - 1 - sensitivity|specificity
- Population with COVID antibodies: 5%
- I tested positive! Do I have COVID antibodies?
- $P(\text{ab}|\text{pos}) = \frac{P(\text{pos}|\text{ab})P(\text{ab})}{P(\text{pos})}$ (Bayes rule)
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos})}$
- Marginal: $P(\text{pos}) = \sum_A P(\text{pos}, A) = P(\text{pos}, \text{ab}) + P(\text{pos}, !\text{ab})$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos},\text{ab}) + P(\text{pos},!\text{ab})}$
- Product rule: $P(\text{pos}, A) = P(\text{pos}|A)\ P(A)$
  - $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{P(\text{pos}|\text{ab})P(\text{ab}) + P(\text{pos}|!\text{ab})P(!\text{ab})}$
- $P(\text{ab}|\text{pos}) = \frac{0.95 \cdot 0.05}{0.95 \cdot 0.05 + 0.05 \cdot 0.95} = 0.5$ (!)

Continuous Random Variable

- $X \in \mathbb{R}$

Continuous Random Variable

- $X \in \mathbb{R}$
- $P(X = x) = 0$ for all $x \in \mathbb{R}$

Continuous Random Variable

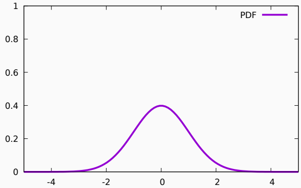
- $X \in \mathbb{R}$
- $P(X = x) = 0$ for all $x \in \mathbb{R}$ (!?)

Continuous Random Variable

- $X \in \mathbb{R}$
- $P(X = x) = 0$ for all $x \in \mathbb{R}$ (!?)

$$P(X \in (a, b)) = \int_a^b f(x)dx$$

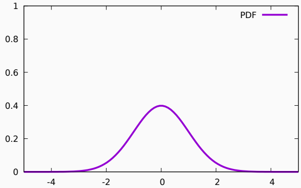
- Probability density function (PDF) $f(x)$, for $x \in \Omega$



Continuous Random Variable

- $X \in \mathbb{R}$
- $P(X = x) = 0$ for all $x \in \mathbb{R}$ (!?)

$$P(X \in (a, b)) = \int_a^b f(x)dx$$



- Probability density function (PDF) $f(x)$, for $x \in \Omega$

$$\int_\Omega f(x)dx = 1 \text{ and } f(x) \geq 0, \text{ for } x \in \Omega$$

Continuous Random Variable

- $X \in \mathbb{R}$
- $P(X = x) = 0$ for all $x \in \mathbb{R}$ (!?)

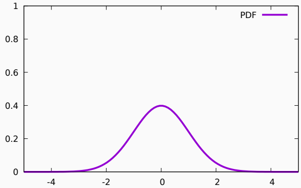
$$P(X \in (a, b)) = \int_a^b f(x)dx$$

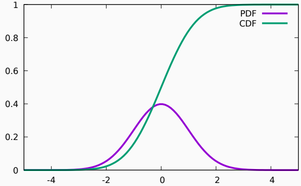
- Probability density function (PDF) $f(x)$, for $x \in \Omega$

$$\int_\Omega f(x)dx = 1 \text{ and } f(x) \geq 0, \text{ for } x \in \Omega$$

- Cumulative distribution function (CDF) $F(x) = P(X \leq x)$

$$F(x) = \int_{-\infty}^x f(t)dt$$





Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$
- Target variables: $Y = (Y_1, \ldots, Y_k)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$
- Target variables: $Y = (Y_1, \ldots, Y_k)$
- Hidden/latent variables: $H = (H_1, \ldots, H_n)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$
- Target variables: $Y = (Y_1, \ldots, Y_k)$
- Hidden/latent variables: $H = (H_1, \ldots, H_n)$
- Want posterior: $P(Y|E)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\; E_1, \ldots, E_m,\; Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\; E_1, \ldots, E_m,\; Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$
- Target variables: $Y = (Y_1, \ldots, Y_k)$
- Hidden/latent variables: $H = (H_1, \ldots, H_n)$
- Want posterior: $P(Y|E)$

$$P(Y|E) = \frac{P(Y,E)}{P(E)} = \frac{\sum_H P(Y,E,H)}{P(E)} \propto \sum_H P(Y,E,H)$$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$
- Target variables: $Y = (Y_1, \ldots, Y_k)$
- Hidden/latent variables: $H = (H_1, \ldots, H_n)$
- Want posterior: $P(Y|E)$

$$P(Y|E) = \frac{P(Y,E)}{P(E)} = \frac{\sum_H P(Y,E,H)}{P(E)} \propto \sum_H P(Y,E,H)$$

- Inference problem: compute $P(Y|E)$

Inference

- Many variables: $X = (H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Joint distribution: $P(H_1, \ldots, H_n,\ E_1, \ldots, E_m,\ Y_1, \ldots, Y_k)$
- Observed variables (evidence): $E = (E_1, \ldots, E_m)$
- Target variables: $Y = (Y_1, \ldots, Y_k)$
- Hidden/latent variables: $H = (H_1, \ldots, H_n)$
- Want posterior: $P(Y|E)$

$$P(Y|E) = \frac{P(Y,E)}{P(E)} = \frac{\sum_H P(Y,E,H)}{P(E)} \propto \sum_H P(Y,E,H)$$

- Inference problem: compute $P(Y|E)$
- Issue: size of table $P(Y_{1:k}, E_{1:m}, H_{1:n})$ is d^{k+m+n}
  - d: number of possible values for each variable
  - If all binary variables: 2^{k+m+n}
  - Remember those $H_{1:n}$?

Cheat Sheet (1)

- Random variable X
  - Values $x \in \text{dom}(x) \rightarrow$ probabilities $P(X = x) \in [0, 1]$
- Probability distribution of X
  - Table (array) of probabilities for each value $x \in \text{dom}(X)$
  - Normalization: $\sum_X P(X) = 1$
- Joint distribution $P(X, Y)$
  - Table (matrix) of probabilities for each value $x, y \in \text{dom}(X) \times \text{dom}(Y)$
- Marginal $P(X) = \sum_Y P(X, Y)$
  - Summing along columns/rows (Y)
- Conditional $P(X|Y) = \frac{P(X,Y)}{P(Y)}$
  - Normalizing each column/row (Y)

Cheat Sheet (2)

- Properties

$$P(X, Y) = P(X|Y)\ P(Y) = P(Y|X)\ P(X)$$

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | X_1, \ldots, X_{i-1})$$

$$P(Y|X) = \frac{P(X|Y)}{P(X)} P(Y) \quad \left[\text{posterior} = \frac{\text{likelihood}}{\text{evidence}}\ \text{prior}\right]$$

- Inference: to compute

$$P(Y_{1:k}|E_{1:m}) = \frac{P(Y_{1:k}, E_{1:m})}{P(E_{1:m})} \propto \sum_{H_{1:n}} P(Y_{1:k}, E_{1:m}, H_{1:n})$$

Supervised Machine Learning Recap

Supervised Machine Learning

- Goal
  - Find $f : X \to Y$
  - Deterministic mapping: $y = f(x)$
  - Set of inputs: X
  - Set of target variables: Y (outputs)

Supervised Machine Learning

- Goal
  - Find $f : X \to Y$
  - Deterministic mapping: $y = f(x)$
  - Set of inputs: X
  - Set of target variables: Y (outputs)
- Relation between X and Y
  - Joint probability distribution: $P(X, Y)$
  - Generally unknown

Supervised Machine Learning (cont.)

- If $P(X, Y)$ was known

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

Supervised Machine Learning (cont.)

- If $P(X, Y)$ was known

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

$$P(y|x) = \frac{P(x, y)}{P(x)} = \frac{P(x, y)}{\sum_{y' \in Y} P(x, y')}$$

Supervised Machine Learning (cont.)

- If $P(X, Y)$ was known

$$P(Y|X) = \frac{P(X, Y)}{P(X)}$$

$$P(y|x) = \frac{P(x, y)}{P(x)} = \frac{P(x, y)}{\sum_{y' \in Y} P(x, y')}$$

$$f(x) = \text{argmax}_{y \in Y} P(y|x)$$

Supervised Machine Learning (cont.)

- $P(X, Y)$: usually not necessary
- Learn a model for $P(Y|X)$ directly

Supervised Machine Learning (cont.)

- $P(X, Y)$: usually not necessary
- Learn a model for $P(Y|X)$ directly
- Or even
  - Learn deterministic $f : X \to Y$ instead of $P(Y|X)$
  - Find $f(x)$ that minimizes generalization error
    - Error for new and unseen $x \in X$

Generalization Error

Generalization error of $f(x)$ (theoretical risk) is the expected loss

$$R[f] = \int_{X \times Y} \ell(x, y, f) dP(X, Y)$$

where $\ell : X \times Y \times F \to \mathbb{R}_0^+$ is a *loss function*

Generalization Error

Generalization error of $f(x)$ (theoretical risk) is the expected loss

$$R[f] = \int_{X \times Y} \ell(x, y, f) dP(X, Y)$$

where $\ell : X \times Y \times F \rightarrow \mathbb{R}_0^+$ is a *loss function*

- BUT: joint probability $P(X, Y)$ is still unknown
- Therefore: $\inf_f R[f]$ cannot be computed directly

Generalization Error

Generalization error of $f(x)$ (theoretical risk) is the expected loss

$$R[f] = \int_{X \times Y} \ell(x, y, f) dP(X, Y)$$

where $\ell : X \times Y \times F \to \mathbb{R}_0^+$ is a *loss function*

- BUT: joint probability $P(X, Y)$ is still unknown
- Therefore: $\inf_f R[f]$ cannot be computed directly
- Solution
  - Approximate $R[f]$ using an N-sample (training set)

    $$\mathcal{D} = \{(x_n, y_n)\}_{n=1 \ldots N} \in X \times Y$$

    drawn *independently and identically distributed* (iid) from $P(X, Y)$

Empirical Risk Minimization

Minimize the empirical risk on $\mathcal{D} = \{(x_n, y_n)\}_{n=1\ldots N}$

$$\hat{R}_{\mathcal{D}}[f] = \frac{1}{N}\sum_{n=1}^{N} \ell(x_n, y_n, f)$$

Asymptotic convergence: $\lim_{N\to\infty} \hat{R}_{\mathcal{D}}[f] = R[f]$

Empirical Risk Minimization

Minimize the empirical risk on $\mathcal{D} = \{(x_n, y_n)\}_{n=1\ldots N}$

$$\hat{R}_{\mathcal{D}}[f] = \frac{1}{N}\sum_{n=1}^{N} \ell(x_n, y_n, f)$$

Asymptotic convergence: $\lim_{N\to\infty} \hat{R}_{\mathcal{D}}[f] = R[f]$

- *Example* regression tasks on the squared loss

$$\ell(x, y, f) = (f(x) - y)^2$$

$$\hat{R}_{\mathcal{D}}[f] = \frac{1}{N}\sum_{n=1}^{N} (f(x_n) - y_n)^2$$

Thank you!

Ulf Brefeld
brefeld@leuphana.de

Eraldo R. Fernandes
eraldo.fernandes@leuphana.de