Data-driven Modeling - Machine Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

Instructor: Heinz Koeppl

LECTURE 2

TECHNISCHE UNIVERSITÄT DARMSTADT

Review of probability theory and statistics

based on Larry Wasserman, All of Statistics - A Concise Course in Statistical Inference, Springer Texts in Statistics, Springer, 2004

MOTIVATION

Recall the general setting of supervised learning from the last lecture:

Given: A labeled set of input-output pairs $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N \in (\mathcal{X} \times \mathcal{Y})^N$

Goal: Learn a mapping $f : \mathcal{X} \to \mathcal{Y}$ from inputs $\mathbf{x} \in \mathcal{X}$ to outputs $y \in \mathcal{Y}$

Problem: Real data is noisy, i.e.

$$y_i = f(\mathbf{x}_i) + \epsilon_i$$

with some unknown perturbation ϵ_i.

How can we include the uncertainty of the data into the learned representation f?

BASIC PROBABILITY

Experiment and outcomes

TECHNISCHE UNIVERSITÄT DARMSTADT

Probability theory is a natural way to model uncertainty.

Consider an *experiment* with a set Ω of possible *outcomes* $\omega \in \Omega$ (also called *sample space*). The outcomes are mutually exclusive and also called *atomic events*.

Example: A fair dice is thrown once.

- The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$

The experimental outcomes of interest are called *Events* (can be atomic events)
Example (Event): A fair dice is thrown once. We are interested whether it is an

even number.

- Event "even number": $A = \{2, 4, 6\} = \{2\} \cup \{4\} \cup \{6\}$

General events $A \subseteq \Omega$ can be constructed from atomic events.

BASIC PROBABILITY

Frequentist probability

We want to assign a number $P(A)$ to each event with

- A never occurs $\quad \Rightarrow \quad P(A) = 0$
- A always occurs $\quad \Rightarrow \quad P(A) = 1$

Definition: If a random experiment is performed n times and the event A occurs n_A times, the probability $P(A)$ is defined as the limit of the *relative frequency*

$$P(A) = \lim_{n\to\infty} \frac{n_A}{n} \,.$$

Proposition: If Ω is finite and all atomic events are equally likely we have

$$P(A) = \frac{|A|}{|\Omega|} \,.$$

The probability of an event is the “volume” of the event compared to the “volume” of the sample space.

BASIC PROBABILITY

Frequentist probability

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: A fair dice is thrown once. What is the probability to get

- an even number?
- a prime number?
- an even prime number?

Answer:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$
- Event "even number": $A_1 = \{2, 4, 6\}$

$$P(A_1) = 3/6 = 0.5$$

- Event "prime number": $A_2 = \{2, 3, 5\}$

$$P(A_2) = 3/6 = 0.5$$

- Event "even prime number": $A_1 \cap A_2 = \{2\}$

$$P(A_1 \cap A_2) = 1/6 \approx 0.17$$

The set Σ of all events of interest is called the *event space*. We saw that for a collection of events

- any *union* of events should also be events
- the complements of events $A^C = \Omega \backslash A$ should also be events

...because we can deduce their probability by logic.

Note that by de Morgan's law $(A \cup B)^C = A^C \cap B^C$, the event space Σ is also closed under intersections.

Formally, this leads to the concept of a σ-algebra.
Definition: Let Ω be a set. A collection of subsets Σ of Ω is called a σ-algebra on Ω if

1. $\Omega \in \Sigma$
2. $A \in \Sigma \quad \Rightarrow \quad A^C \in \Sigma$
3. $A_1, A_2, \ldots \in \Sigma \quad \Rightarrow \quad \bigcup_n A_n \in \Sigma$

The event space Σ needs to be a σ-algebra!

BASIC PROBABILITY

Event space

Example: Toss two coin at once. What are the sample space and the event space?

- The sample space is $\Omega = \{HH, TT, HT, TH\}$.
- Assume we are interested in (non-atomic) events $E_1 = \{HH, TT\}$ and $E_2 = \{HT, TH, TT\}$.
- Then, by the requirements of the $\sigma-$algebra

$$\Sigma = \{\emptyset, \Omega, \{HH, TT\}, \{HT, TH, TT\}, \{HT, TH\}, \{HH\}, \{HT, TH, HH\}, \{TT\}\}$$

Note that for finite Ω, the powerset $\mathscr{P}(\Omega)$ (the set of all subsets) would always give a valid Σ, but not the smallest possible one (i.e., $|\mathscr{P}(\Omega)| = 2^4 = 16$).

Generally, let $\mathcal{E}$ be an arbitrary collection of events $E_i \subset \Omega$ of interest. Then we say that $\sigma(\mathcal{E})$ is the σ-algebra generated by $\mathcal{E}$.

Uncountable sample space

TECHNISCHE UNIVERSITÄT DARMSTADT

The event space Σ is especially important if the sample space is uncountable (e.g. $\Omega = \mathbb{R}$). Then every atomic outcome $\omega \in \Omega$ has probability / measure zero (or strictly, is undefined).

For instance, asking what is the probability for drawing a certain random number x from a Gaussian distribution does not make sense.

The σ-algebra generated by all open intervals $\{(a, b) \mid a < b, a \in \mathbb{R}, b \in \mathbb{R}\}$ of $\Omega = \mathbb{R}$ is called the Borel σ-algebra.

As a consequence all half-open, closed intervals and combination thereof will be part of the Borel σ-algebra, e.g. $(-\infty, a] \cup [b, \infty)$, $(-\infty, a]$, $[b, \infty), \ldots$.

Hence, asking for the probability of drawing a $x \in (a, b)$ from a Gaussian distribution makes sense.

Mathematical probability is based on an axiomatic formulation.

Definition: Let Ω be a sample space and Σ be a σ-algebra on Ω. A *probability function* is a map $P : \Sigma \to [0, 1]$ with

1. $P(A) \geq 0$ for all $A \in \Sigma$,
2. $P(\Omega) = 1$,
3. Any countable sequence of disjoint events A_i satisfies

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i) \, .$$

Definition: A *probability space* is a triple (Ω, Σ, P) where

1. Ω is the set of all possible outcomes (sample space),
2. Σ a σ-algebra on Ω called the event space,
3. P is a probability function on Σ.

BASIC PROBABILITY

Joint probability

If some events A and B are not disjoint, we can use the axioms above to get

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)\,.$$

Definition: $P(A \cap B)$ is called the *joint probability* of the events A and B.

Example: A fair dice is thrown once. What is the probability to get an even prime number?

- even numbers: $A = \{2, 4, 6\}$
- prime numbers: $B = \{2, 3, 5\}$
- even prime numbers: $A \cap B = \{2\}$
- even or prime number $A \cup B = \{2, 3, 4, 5, 6\}$

$$\Rightarrow \quad \begin{cases} P(A) = P(B) = \frac{1}{2} \\ P(A \cap B) = \frac{1}{6} \\ P(A \cup B) = \frac{5}{6} \end{cases}$$

The results above are related via $P(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{5}{6} = \frac{1}{6}$.

BASIC PROBABILITY

Conditional probability

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: For $P(B) > 0$ the probability of the event A given that event B has occurred is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \,.$$

Example: A fair dice is thrown once. What is the probability to get an even number given that the result is a prime number?

- $P(A \cap B) = 1/6$
- and $P(B) = 1/2$ $\quad \Rightarrow \quad P(A \mid B) = \dfrac{1/6}{1/2} = \dfrac{1}{3}$

Remarks:

1. The equation above is often stated in a form known as the *product rule of probability*: $P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A)$.
2. The conditional probability satisfies the *axioms of probability* and can thus be seen as a probability function on the reduced sample space B.

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: Two events A and B are called *independent events* if

$$P(A \cap B) = P(A) \cdot P(B)\,.$$

As a consequence, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$.

A similar statement holds for conditional probabilities.

Definition: Two events A and B are conditionally independent given C if

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)\,.$$

Remarks:

1. Conditional independence does not imply independence or vice versa.
2. We will come back to conditional independence in the section on *probabilistic graphical models*.

BASIC PROBABILITY

Law of total probability

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: A *partition* $\{A_i : i = 1, 2, \ldots\}$ of a set Ω is a non-empty collection of pairwise disjoint subsets $A_i \subset \Omega$ such that $\bigcup_i A_i = \Omega$.

Proposition: For a partition $\{A_i : i = 1, 2, \ldots\}$ and an arbitrary event $B \subset \Omega$

$$P(B) = \sum_i P(B \cap A_i)$$

or equivalently using the product rule

$$P(B) = \sum_i P(B \mid A_i) P(A_i) \, .$$

Remarks:

1. The result is also known as the *sum rule of probability*.
2. $P(B \cap A_i)$ can be understood as a joint probability of B and A_i. $P(B)$ is then called the *marginal probability* of the event B.

BASIC PROBABILITY

Bayes theorem

Theorem: For two events A and B with $P(A) > 0$ and $P(B) > 0$, the conditional probabilities are related via

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \, .$$

A more general form of Bayes theorem considers an event B and a partition $\{A_i : i = 1, 2, \ldots .\}$. Using the law of total probability, one gets

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{\sum_j P(B \mid A_j)P(A_j)} \, .$$

Proof of the basic form: Using the definition of conditional probability and the product rule yields

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B \mid A)P(A)}{P(B)} \, .$$

BASIC PROBABILITY

Bayes theorem

TECHNISCHE UNIVERSITÄT DARMSTADT

Example: There are three email categories with prior probabilities:

- $A_1 =$ "spam", $\quad P(A_1) = 0.7$
- $A_2 =$ "low priority", $\quad P(A_2) = 0.2$
- $A_3 =$ "high priority", $\quad P(A_3) = 0.1$

$$P(A_1) + P(A_2) + P(A_3) = 1$$

Let B denote the event that an email contains the word "free". The conditional probabilities that an email contains the word "free" given the category are $P(B \mid A_1) = 0.9$, $P(B \mid A_2) = 0.01$, $P(B \mid A_3) = 0.01$. When receiving an email with the word "free", what is the probability that it is spam?

$$\begin{aligned} P(A_1 \mid B) &= \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} \\ &= \frac{0.9 \cdot 0.7}{0.9 \cdot 0.7 + 0.01 \cdot 0.2 + 0.01 \cdot 0.1} = 0.995 \end{aligned}$$

RANDOM VARIABLES

Definition

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: A *random variable* (RV) is a function $X : \Omega \to \mathcal{X}$ that assigns an element of $\mathcal{X}$ to each outcome $\omega \in \Omega$.

Remark: Typically, we consider

- *discrete random variables* with $\mathcal{X} = \mathbb{N}$ or
- *continuous random variables* with $\mathcal{X} = \mathbb{R}$.

Notation:

- X denotes a random variable (i.e. a function)
- x denotes a particular realization of X (usually a number)
- $X(\omega) = x$ means that the random variable X takes the particular value x
- $\{X \leq x\} = \{\omega : X(\omega) \leq x\}$ is the set of all outcomes $\omega \in \Omega$ for which $X(\omega)$ takes values less than or equal to x

RANDOM VARIABLES

Distribution function

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: Let $\mathcal{X} = \mathbb{N}, \mathbb{R}$. The function $F_X(x) := P(X \leq x)$ is called the *cumulative distribution function* (CDF).

Properties of the CDF

1. $0 \leq F_X(x) \leq 1$ with $F_X(-\infty) = 0$ and $F_X(+\infty) = 1$
2. $F_X(x)$ is continuous from the right, i.e. $\lim_{\epsilon \to 0} F_X(x + \epsilon) = F_X(x)$
3. $F_X(x)$ is non-decreasing, i.e. $F_X(x_1) \leq F_X(x_2)$ for all $x_1 < x_2$
4. For an interval $[a, b]$ we have

$$P(a \leq X \leq b) = F_X(b) - F_X(a)$$

ta errado isso não? -> era pra ser

a < X

Notation: If X follows a particular distribution F we write $X \sim F$.

RANDOM VARIABLES

Probability mass and density function

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: For a discrete random variable X we define the *probability mass function* (PMF) of X by $f_X(x) \equiv P(X = x)$.

Remark: For a continuous random variable, $P(X = x) = 0$ for all x.

Idea: Consider small interval $[x, x + dx]$. Then
$P(x \leq X \leq x + dx) = F_X(x + dx) - F_X(x)$.

Definition: If $F_X(x)$ is differentiable, the *probability density function* (PDF) is defined as

$$f_X(x) \equiv \frac{dF_X(x)}{dx} \quad \text{and hence} \quad P(x \leq X \leq x + dx) = f_X(x)dx\,.$$

Properties:

1. $F_X(x) = \int_{-\infty}^{x} f_X(x')dx'$
2. $P(a < X < b) = \int_a^b f_X(x')dx'$

Bernoulli distribution

Let X be a binary random variable, i.e. $\mathcal{X} = \{0, 1\}$ with $P(X = 1) = p$ and $P(X = 0) = 1 - p$ for a parameter $p \in [0, 1]$. Then the PMF can be written as

$$f(x) = p^x (1-p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

and we write $X \sim \mathrm{Bernoulli}(p)$.

Binomial distribution

Assume we draw n samples from a Bernoulli distribution with parameter $p \in [0, 1]$. Let X represent the number of successes ($\equiv$ number of ones). Then X has the PMF

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

and we write $X \sim \mathrm{Binomial}(n, p)$.

RANDOM VARIABLES

Important discrete random variables

TECHNISCHE UNIVERSITÄT DARMSTADT

Poisson distribution

Let $\mathcal{X} = \mathbb{N}_0$ and X a random variable with PMF

$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

where $\lambda \in [0, \infty)$. We say that X has a Poisson distribution with parameter λ and we write $X \sim \mathrm{Poisson}(\lambda)$.

Applications:

- Bernoulli random variables can be used to model a noisy channel that transmits a binary signal.
- Binomial distributions appear in many contexts where *summary statistics* of more complicated distributions are considered.
- Poisson distributions are used to model event counts, e.g. the number of accesses to a server.

RANDOM VARIABLES

Important continuous random variables

TECHNISCHE UNIVERSITÄT DARMSTADT

Uniform distribution

A random variable X has a (continuous) uniform distribution on the interval $[a, b]$, written as $X \sim \mathcal{U}(a, b)$, if it has the PDF

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } x \in [a, b] \\ 0 & \text{otherwise} \end{cases} .$$

Normal distribution

A random variable X has a normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma > 0$, written as $X \sim \mathcal{N}(\mu, \sigma^2)$, if it has the PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) .$$

The normal distribution is very important because

- many quantities can be approximated by a normal distribution,
- it has convenient mathematical properties.

Exponential distribution

A random variable X has an exponential distribution with parameter $\lambda > 0$, written as $X \sim \mathrm{Exp}(\lambda)$, if it has the PDF

$$f(x) = \lambda e^{-\lambda x} \quad \text{for} \quad x > 0\,.$$

Gamma distribution

A random variable X has a Gamma distribution with parameters $\alpha, \beta > 0$, written as $X \sim \Gamma(\alpha, \beta)$, if it has the PDF

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad \text{for } x > 0 \text{ with } \Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy\,.$$

- Exponential distributions describe the waiting time for a memoryless process (e.g. the interarrival times between independent accesses to a server).
- Gamma distributions allow flexible modeling of positive continuous observables by varying the parameters α and β.

MULTIPLE RANDOM VARIABLES

Joints

Definition: For n random variables $X_1, X_2, \ldots, X_n$ the function $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n)$ is called the *joint distribution function*.

Definition: If $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ is differentiable, the *joint density function* is defined as $f_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots, x_n)$

$$f_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots, x_n) = \frac{\partial^n F_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \ldots \partial x_n}$$

Properties:

1. $f_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots, x_n) \geq 0$ for all $(x_1, x_2, \ldots, x_n)$,
2. For any set $A \subset \mathbb{R}^n$ we have

$$P((x_1, x_2, \ldots, x_n) \in A) = \int_A f_{X_1,X_2,\ldots X_n}(x_1, x_2, \ldots, x_n)\, dx_1 dx_2 \ldots dx_n\,.$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: Let $F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)$ denote the joint distribution of $X_1, X_2, \ldots, X_n$. The marginal distribution function of X_i is given by

$$F_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

In the continuous case we obtain the marginal density function by

$$f_{X_i}(x_i) = \int_{\mathbb{R}^{n-1}} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n)\, dx_1 \ldots dx_{i-1} dx_{i+1} \ldots dx_n$$

Remarks:

- Marginals can be defined for any subset of the RV's $X_1, X_2, \ldots, X_n$.
- The process of calculating marginals is called *marginalization*.

MULTIPLE RANDOM VARIABLES

Conditionals

For simplicity, consider two random variables X_1 and X_2 with a joint distribution F_{X_1,X_2}. We are interested in the distribution of X_1 for a given value of X_2.

Definition: For X_1, X_2 discrete and $f_{X_2}(x_2) > 0$, the *conditional probability mass function* is

$$f_{X_1|X_2}(x_1|x_2) := P(X_1 = x_1|X_2 = x_2) = \frac{P(X_1 = x_1, X_2 = x_2)}{P(X_2 = x_2)} = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \,.$$

Definition: For X_1, X_2 continuous and $f_{X_2}(x_2) > 0$, the *conditional probability density function* is[1]

$$f_{X_1|X_2}(x_1|x_2) := \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad \text{and} \quad P(a_1 \leq X_1 \leq b_1|X_2 = x_2) = \int_{a_1}^{b_1} f_{X_1|X_2}(x_1|x_2)\, dx_1 \,.$$

[1] Note that we are conditioning on the event $X_2 = x_2$ which has probability zero. A rigorous treatment of conditional random variables requires a measure theoretic approach.

MULTIPLE RANDOM VARIABLES

Independence

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: The random variables $X_1, X_2, \dots, X_n$ are said to be *independent* if for every $A_1, A_2, \dots, A_n$

$$P(X_1 \in A_1, X_2 \in A_2, \dots, X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i) \quad \Leftrightarrow \quad f(x_1, x_2, \dots x_n) = \prod_{i=1}^{n} f_{X_i}(x_i)$$

where f is the PMF in the discrete case and the PDF in the continuous case.

Definition: If $X_1, X_2, \dots, X_n$ are independent and all X_i have the same marginal distribution F, we say that $X_1, X_2, \dots X_n$ are *independent and identically distributed* (i.i.d.).

Remark: We can also see the $X_1, X_2, \dots X_n$ IID as a sample of size n from the distribution F. We come back to this idea when we discuss statistical inference.

TECHNISCHE UNIVERSITÄT DARMSTADT

Multivariate normal distribution

Let $X = (X_1, X_2, \ldots, X_n)$ be a vector valued RV on $\mathbb{R}^n$. A RV X has a *multivariate normal distribution* with parameters $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n \times n}$ symmetric positive definite, if it has the PDF

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

- We write $X \sim \mathcal{N}(\mu, \Sigma)$.
- $|\Sigma| := |\det \Sigma|$ denotes the absolute value of the determinant of Σ.
- A matrix Σ is symmetric if $\Sigma^T = \Sigma$.
- A symmetric matrix Σ is positive definite if $x^T \Sigma x > 0$ for all nonzero vectors x.
- Σ is called the covariance matrix.
- Σ^{-1} is the inverse of Σ and is called the precision matrix Λ.

In applications, the full distribution of a random variable X is usually inaccessible. We therefore consider certain summary functions.

Definition: The *expected value* of a discrete RV X is defined as

$$\mathrm{E}[X] = \sum_{x \in \mathcal{X}} x\, f_X(x)$$

where $f_X(x)$ is the PMF of X.

Definition: The *expected value* of a continuous RV X is defined as

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx$$

where $f_X(x)$ is the PDF of X.

OPERATIONS ON RANDOM VARIABLES

Properties of the expectation

TECHNISCHE UNIVERSITÄT DARMSTADT

Remarks:

- $\mathrm{E}[X]$ is also called the *mean* or *first moment* of X.
- Generalization to multiple random variables is straightforward.
- For an RV X with density f_X and a function g, define the new RV $Y = g(X)$. Then

$$\mathrm{E}[Y] := \mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x)\, f_X(x)\, dx$$

Important properties: Let X, Y be general RVs with $\mathrm{E}[X], \mathrm{E}[Y] < \infty$

1. Linearity: For RVs X, Y and constants α, β: $\mathrm{E}[\alpha X + \beta Y] = \alpha \mathrm{E}[X] + \beta \mathrm{E}[Y]$
2. Monotonicity: If $X \leq Y$ ($F_X(x) \leq F_Y(x)$, $\forall x$), then also $\mathrm{E}[X] \leq \mathrm{E}[Y]$
3. For X and Y independent: $\mathrm{E}[X \cdot Y] = \mathrm{E}[X] \cdot \mathrm{E}[Y]$

OPERATIONS ON RANDOM VARIABLES

Expectation of the normal distribution

Example: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then

$$
\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} x \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] dx \qquad \Big| \, z := x - \mu, \quad dz = dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (z+\mu) \exp\left[-\frac{z^2}{2\sigma^2}\right] dz \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} z \exp\left[-\frac{z^2}{2\sigma^2}\right] dz + \frac{\mu}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left[-\frac{z^2}{2\sigma^2}\right] dz \\
&= 0 + \mu \\
&= \mu
\end{aligned}
$$

OPERATIONS ON RANDOM VARIABLES

Higher moments and variance

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: For the RV X set $g(X) = X^n$. The respective expectation $\mathrm{E}[X^n]$ is called the n-th order moment or simply the n-th moment.

Definition: For a *RV* X with $\mathrm{E}[X], \mathrm{E}[X^2] < \infty$ the variance is defined as $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$.

- The variance can also be written as $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$.
- It is a measure of the spread of a distribution around its mean.
- The standard deviation is related to the variance via $\mathrm{std}[X] = \sqrt{\mathrm{Var}[X]}$.

Example: Let $X \sim \mathcal{N}(\mu, \sigma^2)$. We already computed $\mathrm{E}[X] = \mu$. For the variance, we get

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[(X-\mu)^2] \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} (x-\mu)^2 \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right] dx \qquad \Big| \; z := x - \mu, \quad dz = dx \\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} z \cdot z \exp\left[-\frac{z^2}{2\sigma^2}\right] dz \qquad \Big| \; \text{integration by parts with } u \cdot v' \\
&= -\frac{1}{\sqrt{2\pi}\sigma} z \cdot \sigma^2 \exp\left[-\frac{z^2}{2\sigma^2}\right] \Bigg|_{-\infty}^{\infty} + \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} 1 \cdot \sigma^2 \exp\left[-\frac{z^2}{2\sigma^2}\right] dz \\
&= 0 + \sigma^2 \\
&= \sigma^2
\end{aligned}
$$

OPERATIONS ON RANDOM VARIABLES

Covariance of multiple random variables

TECHNISCHE UNIVERSITÄT DARMSTADT

Definition: For two random variables X_1, X_2, we define the *covariance*

$$\mathrm{Cov}[X_1, X_2] = \mathsf{E}[(X_1 - \mathsf{E}[X_1])(X_2 - \mathsf{E}[X_2])]$$

and the *correlation*

$$\mathrm{Corr}[X_1, X_2] = \frac{\mathrm{Cov}[X_1, X_2]}{\sqrt{\mathrm{Var}[X_1]\mathrm{Var}[X_2]}} .$$

Remarks:

- The definition extends to multiple RVs by calculating the covariances and correlations for all pairs.
- The correlation is $+1$ in case of a perfect increasing linear relationship and -1 in case of a perfect decreasing linear relationship.

STATISTICS

Statistical inference

So far, we have looked at random variables following certain types of distribution. Now consider the *inverse problem*:

Given $X_1, \dots X_n \sim F$ i.i.d., how can we learn (some properties of) F?

- This task is known as *statistical inference* or *learning*.
- Statistics is deeply connected with machine learning.

To obtain a solvable problem, we need to restrict the class of candidates F, e.g. by

- choosing a known family F_θ defined by some parameter vector θ
- imposing constraints on the shape of F (smoothness in kernel-based methods)
- imposing constraints on the structure of F (factorization in variational inference)

??

Definition: Assume we have $X_1, \ldots, X_n$ i.i.d. samples from F_θ. A *point estimator* $\hat{\theta}_n$ of θ is a function

$$\hat{\theta}_n = g(X_1, \ldots, X_n)\,.$$

We call $\hat{\theta}_n$ *unbiased* if

$$\mathrm{E}[\hat{\theta}_n] = \theta$$

and *consistent* if

$$\hat{\theta}_n \longrightarrow \theta \quad \text{for } n \to \infty.$$

Two important estimators are the *sample mean* $\bar{X}_n$ and the *sample variance* S_n^2

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad , \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\,.$$

TECHNISCHE UNIVERSITÄT DARMSTADT

The behavior of the sample mean for a large number of samples is described by two important theorems. Let $\mu = \mathrm{E}[X]$ denote the expected value of X with $X \sim F$.

Weak law of large numbers (WLLN): If $X_1, \ldots, X_n$ i.i.d. from F, then

$$\bar{X}_n \longrightarrow \mu \quad \text{for } n \to \infty.$$

Central limit theorem (CLT): If $X_1, \ldots, X_n$ i.i.d. from F, then

$$Z_n = \frac{\bar{X}_n - \mu}{\sqrt{\mathrm{Var}[\bar{X}_n]}} \longrightarrow Z \sim \mathcal{N}(0, 1) \quad \text{for } n \to \infty.$$

Remarks:

- The sample mean is a consistent estimator of the true mean.
- The CLT states that the sample mean for a large number n of samples is approximately normally distributed.

Idea: Given samples $X \sim F_\theta$ (or from density f_θ) can we determine the most likely θ that gave rise to the samples?

Definition: Let $X_1, \ldots, X_n$ i.i.d. be continuous random variables with PDF $f_\theta(x_i)$. The *likelihood function* $L_n(\theta)$ is defined as the joint

$$L_n(\theta) \equiv f_\theta(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_\theta(x_i)\,.$$

Definition: The *maximum likelihood estimator* (MLE) is defined as the value of θ that maximizes the likelihood, i.e.

$$\hat{\theta}_n = \arg\max_\theta L_n(\theta) = \arg\max_\theta \log L_n(\theta)\,.$$

In practice, it is often more convenient to maximize the logarithm of the likelihood instead.

Example: Let $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with σ known. What is the MLE of μ?

Likelihood: $L_n(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i - \mu)^2\right]$

Log-likelihood: $\log L_n(\mu) = -n \log \sqrt{2\pi}\sigma - \frac{1}{2\sigma^2} \sum_{i=1}^{n}(x_i - \mu)^2$

- The log-likelihood is a differentiable function of the parameter μ.
- We can find the MLE by solving $\frac{d}{d\mu} \log L_n(\mu) \stackrel{!}{=} 0$.

$\Rightarrow$ MLE: $\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i$

Remark: For i.i.d. Gaussian RVs the maximum likelihood method recovers the sample mean estimator.

STATISTICS

Bayesian inference

Idea: We treat the unkown model parameter θ as a random variable encoding our epistemic uncertainty (our belief).

Assume we have some prior information on the parameter θ before collecting samples $X_1, \dots, X_n$. We would like to update the belief about θ with the new information provided by the samples.

Bayesian solution: The previous information is encoded in a *prior probability* distribution $f(\theta)$. The joint distribution over parameters and samples is given by

$$f(x_1, \dots, x_n, \theta) = f(x_1, \dots, x_n \mid \theta) f(\theta) \,,$$

where $f(\cdot \mid \theta) := f_\theta(\cdot)$ is the likelihood function. From Bayes theorem, the *posterior probability* of the parameters given the data is

$$f(\theta \mid x_1, \dots, x_n) = \frac{f(x_1, \dots, x_n \mid \theta) f(\theta)}{f(x_1, \dots, x_n)}$$

where $f(x_1, \dots, x_n) = \int f(x_1, \dots, x_n \mid \theta) f(\theta) d\theta$ is called the *evidence*.