Advanced Machine Learning

R. Vosshall, C. Bogoclu & N. Friedlich

Aug 17, 2021

CONTENTS

Note: This book is currently under development.

Part I

Preface

CHAPTER

ONE

PREFACE

CHAPTER

TWO

NOTATION

Part II

Fundamentals

CHAPTER

THREE

FUNDAMENTALS OF PROBABILITY THEORY

In the present section, we define the basic terms of probability theory and statistics. Moreover, we state the most common examples of discrete and continuous probability distributions.

The content follows the textbooks

"Statistik für Ingenieure - Wahrscheinlichkeitsrechnung und Datenauswertung endlich verständlich"

by Aeneas Rooch and

"Grundlagen der Wahrscheinlichkeitsrechnung und Statistik - Eine Einführung für Studierende der Informatik, der Ingenieur- und Wirtschaftswissenschaften"

by Erhard Cramer and Udo Kamps.

The goal is to avoid unnecessarily complex mathematical backround, but to provide the required framework to understand the subsequent machine learning methods. Nevertheless, for the the sake of completeness, additional references are given from time to time. A more profound mathematical theory can for example be found in "Wahrscheinlichkeitstheorie" by Achim Klenke.

Note: All three books are available free of charge via DigiBib.

3.1 Probability Spaces

Definition

In order to model the outcome of a random experiment, we denote by Ω the **sample space** of all possible outcomes, i.e.,

$$\Omega = \{\omega \mid \omega \text{ is a possible outcome of the random experiment}\}.$$

Accordingly, each element $\omega \in \Omega$ is called an **outcome**. A subset A of Ω of possible outcomes is called an **event**. If A contains only a single outcome ω, i.e., $A = \{\omega\}$ for some $\omega \in \Omega$, A is also called an elementary event.

If we model the rolling of an ordinary cubic dice, the sample space

$$\Omega = \{1, 2, 3, 4, 5, 6\} \tag{1}$$

is given by the 6 possible outcomes. The event A of rolling an even number is given by $A = \{2, 4, 6\} \subset \Omega$ and the elementary event of rolling a six is given by $A = \{6\}$.

3.1.1 Discrete Probability Spaces

Definition

Let Ω be a *finite or countable* sample space and denote by $\mathcal{P}(\Omega) = \{A \mid A \subset \Omega\}$ the set of all subsets of Ω (the so-called power set). Moreover, let $p : \Omega \to [0, 1]$ be a map such that $\sum_{\omega \in \Omega} p(\omega) = 1$. Then, the map $P : \mathcal{P}(\Omega) \to [0, 1]$ given by

$$P(A) := \sum_{\omega \in A} p(\omega) \quad \text{for } A \in \mathcal{P}(\Omega)$$

is called a **discrete probability measure** or a **discrete probability distribution**. The triple $(\Omega, \mathcal{P}(\Omega), P)$ or briefly (Ω, P) is called a **discrete probability space**.

- A probability measure P assigns to each possible event a probability between 0 ("impossible") to 1 ("sure").
- P is completely characterized by the elementary probabilities (i.e., the probabilities of elementary events specified by p) in the case of discrete probability distributions (by definition).
- The condition $\sum_{\omega \in \Omega} p(\omega) = 1$ guarantees that $P(\Omega) = 1$. In other words, it has to be sure that the outcome of a random experiment is indeed in Ω and moreover, $P(\Omega) > 1$ would make no sense in terms of probabilities.

Assumed that we are dealing with a fair dice as in *(1.1)*, it is reasonable to define $p(\omega) := \frac{1}{6}$ for each $\omega = 1, \dots, 6$. Hence, each outcome of a dice roll is each likely. Consequently, the probability of rolling an even number is

$$P(\{2, 4, 6\}) = \sum_{\omega \in \{2,4,6\}} p(\omega) = 3 \cdot \frac{1}{6} = 0.5$$

as expected.

Corollary

As a direct consequence of Definition 1.3, a discrete probability measure has the following properties:

- $0 \leq P(A) \leq 1$ for each event $A \in \mathcal{P}(\Omega)$,
- $P(\Omega) = 1$,
- P is σ-additive, i.e., for pairwise disjoint events A_i, $i \in \mathbb{N}$, it holds

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i).$$

- The statements in *the Corollary* are also called **Kolmogorov axioms**.
- The term "pairwise disjoint" means that two arbitrary events do not have any common elements. For example, the events $\{2, 4, 6\}$ and $\{1, 3\}$ are disjoint, but the events $\{2, 4, 6\}$ and $\{2, 3\}$ are not, since the share the outcome 2.
- The last statement of Corollary 1.6 also holds true for a *finite number* of sets A_i, $i = 1, \dots, n$, by simply choosingchoosing $A_i = \emptyset$ (empty set) for $i > n$. If we consider only two disjoint sets A_1 and A_2, it follows that $P(A_1 \cup A_2) = P(A_1) + P(A_2)$. This means that the probability that the event A_1 or the event A_2 occurs equals the sum of the probabilities, which is intuitive.

3.2 Continuous Probability Spaces

It turns out that the definition of probability spaces requires a different approach in the case of sample spaces that contain *uncountably* many outcomes. For example, the sample space could be given by the real numbers ($\Omega = \mathbb{R}$) or a higher dimensional space (e.g. $\Omega = \mathbb{R}^d$, $d >= 2$). Indeed, the definition of (probability) measures on arbitrary sample spaces turns out to be a complex mathematical problem which is the foundation of **measure theory**. This theory introduces so-called σ-algebras which specify the measurable events, i.e., the events for which it is possible to assign a probability without generating any inconsistencies. An introduction can be found in the first chapter of "Wahrscheinlichkeitstheorie" by Achim Klenke. Measure theory is the foundation of very powerful results, since it enables mathematicians to define probability measures even on infinite dimensional sample spaces such as spaces of functions which lead to so-called stochastic processes. A special case are **Gaussian processes** which turn out to be very useful in the context of machine learning and are an essential part of this lecture.

Luckily, we do not necessarily need to consider measure theory in detail for our purposes. For the mentioned cases ($\Omega = [0, 1]$ or $\Omega = \mathbb{R}^d$, $d >= 1$), we can use ordinary integrals in order to define probabilities at least on "nice" events.

Remark

- Set $C := \{[a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d] \mid -\infty \leq a_i \leq b_i \leq \infty,\ i = 1, \dots, d\}$. An event $A \in C$ is simply a box. For d=1 we obtain an interval $A = [a_1, b_1]$ and for d=2 we get a rectangle $A = [a_1, b_1] \times [a_2, b_2] \subset \mathbb{R}^2$.
- In measury theory, C is a so-called generating system of the **Borel σ-algebra** $\mathcal{B}(\mathbb{R}^d)$. The Borel σ-algebra is the smallest collection of events with sufficiently nice properties which contains alle these boxes.
- $\mathcal{B}(\mathbb{R}^d)$ is fairly abstract. Just rembember that
  - $\mathcal{B}(\mathbb{R}^d)$ contains all events we would like / are able to assign a probability to,
  - there are subsets of $\mathbb{R}^d$ which are not in $\mathcal{B}(\mathbb{R}^d)$, but we do not care, since they are not important.

Definition

Let $\Omega = \mathbb{R}^d$, $d \geq 1$, be the sample space and $f : \mathbb{R}^d \to \mathbb{R}$ be an integrable non-negative function such that

$$\int_{\mathbb{R}^d} f(x)\, dx = 1.$$

Then, the map $P : C \to [0, 1]$ defined by

$$P(A) := \int_{a_d}^{b_d} \cdots \int_{a_1}^{b_1} f(x)\, dx \quad \text{for } A = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d] \in C$$

extends uniquely to $\mathcal{B}(\Omega)$ (not part of the lecture) and this extension is called a **continuous probability measure** or a **continuous probability distribution**. f is called the **probability density function** (PDF) of P or briefly probabilty density or simply density. The triple $(\Omega, \mathcal{B}(\Omega), P)$ or briefly (Ω, P) is called a **continuous probability space**. Furthermore, the function $F_X : \mathbb{R}^d \to [0, 1]$ defined by

$$F(x) := P((-\infty, x_1] \times \cdots \times (-\infty, x_d]) \quad \text{for } x = (x_1, \dots, x_d) \in \mathbb{R}^d$$

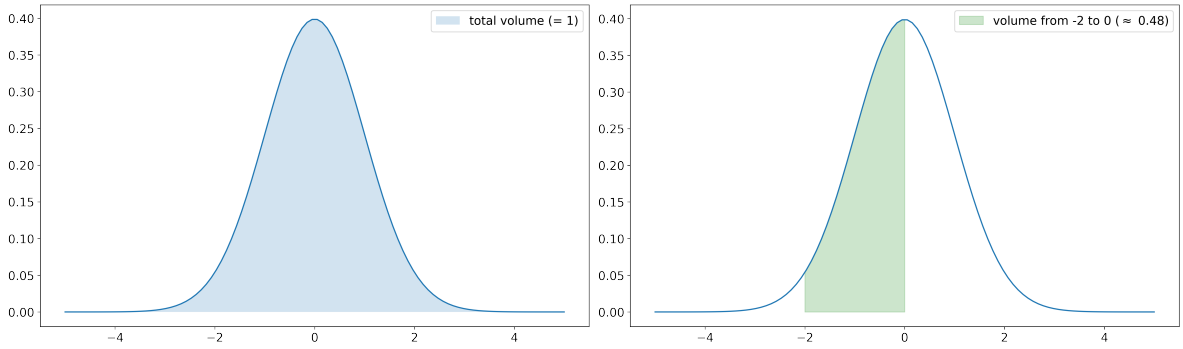
is called the **cumulative distribution function** (CDF) of P.

It holds

$$\int_{\mathbb{R}} \exp\big(-\frac{1}{2}x^2\big)\, dx = \sqrt{2\pi}.$$

Therefore, $f : \mathbb{R} \to \mathbb{R}$ defined by $f(x) := \frac{1}{\sqrt{2\pi}} \exp\big(-\frac{1}{2}x^2\big)$ for $x \in \mathbb{R}$ defines a continuous probability distribution with density f which is called **standard normal distribution**.

- As in the case of discrete probability spaces, a probability measure as defined in Definition 1.9 fulfills the **Kolmogorov axioms** stated in Corollary 1.6.
- In order to unify the notation of discrete and continuous probability spaces, we denote a general probability space by $(\Omega, \mathcal{F}, P)$, where $\mathcal{F}$ denotes a σ-algebra (in our case either $\mathcal{P}(\Omega)$ or $\mathcal{B}(\Omega)$).
- Keep in mind that f is simply a non-negative function whose volume under the graph is exactly one and the probability of some event A is the volume under the graph of f restricted to A. In the plot below, $P([-2, 0]) \approx 0.48$ is illustrated for a standard normal distribution. In other words, the probability to observe an outcome between -2 and 0 in a standard normally distributed experiment is approximately 48%.



3.3 Random Variables

Imagine that we perform multiple independent random experiments by rolling repeatedly (n-times) a fair dice as in Example 1.5. The corresponding sample space is given by

$$\Omega = \{\omega = (\omega_1, \omega_2, \dots, \omega_n) \mid \omega_i \in \{1, 2, 3, 4, 5, 6\} \text{ for } i = 1, \dots, n\}.$$

Since the experiments are independent and we consider a fair dice, it is reasonable to define

$$p(\omega) = \frac{1}{6^n} \quad \text{for each } \omega \in \Omega$$

which results in a discrete probability space $(\Omega, \mathcal{P}(\Omega), P)$. Eventually, we are not interested in events with respect to Ω, but for example in the average outcome of the experiments or the number of times of rolling a six. Instead of modelling these experiments directly by redefining Ω and p, it is very useful to apply the concept of random variables:

Definition

Let $(\Omega, \mathcal{F}, P)$ be a probabiliy space. A map $X : \Omega \to \mathbb{R}^d$, $d \geq 1$, is called a real-valued **random variable**, if

$$X^{-1}(A) := \{\omega \in \Omega \mid X(\omega) \in A\} \in \mathcal{F}$$

for each $A \in \mathcal{B}(\mathbb{R}^d)$ and the probability measure $P_X : \mathcal{B}(\Omega) \to [0, 1]$ given by

$$P_X(A) := P(X^{-1}(A)) \quad \text{for } A \in \mathcal{B}(\Omega)$$

is called the **distribution of** X **under** P. If P_X admits a probability density as defined in Definition 1.9, then we denote the density by f_X. Furthermore, the cumulative distribution function of P_X is denoted by F_X.

- If $(\Omega, \mathcal{F}, P)$ is a discrete probability space as defined in Definition 1.3, **each** map $X : \Omega \to \mathbb{R}^d$ is a random variable, since $\mathcal{F} = \mathcal{P}(\Omega)$ contains all subsets of Ω.

- If $(\Omega, \mathcal{F}, P)$ is a continuous probability space as defined in Definition 1.9, it can be shown that at least each **continuous** map $X : \Omega \to \mathbb{R}^d$ is a random variable.
- If $d > 1$, $X : \Omega \to \mathbb{R}^d$ is a random variable if and only if each component $X_i : \Omega \to \mathbb{R}$ is a random variable.

3.4 Independence

Definition

Let $(\Omega, \mathcal{F}, P)$ be a probability space. Then two events $A, B \in \mathcal{F}$ are called **independent**, if

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

Let $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$ be two random variables. Then X and Y are called **independent random variables**, if the events $X^{-1}(A)$ and $Y^{-1}(B)$ are independent for all $A, B \in \mathcal{B}(\mathbb{R})$. This is equivalent to the property

$$F_{(X,Y)}(x, y) = F_X(x)F_Y(y) \quad \text{for all } x, y \in \mathbb{R}$$

and if the corresponding densities exist to

$$f_{(X,Y)}(x, y) = f_X(x)f_Y(y) \quad \text{for all } x, y \in \mathbb{R}.$$

Independence basically means that the occurrence of event A has no impact on the occurrence of event B or in terms of random variables, the random variables should not impact each other. The definition of independent random variables generalizes easily to $d > 1$.

3.5 Important Probability Distributions

3.5.1 Discrete Distributions

3.5.2 Continuous Distributions

CHAPTER

FOUR

BAYESIAN VS. FREQUENTISTS VIEW

4.1 Bayes' Theorem

CHAPTER

FIVE

MLE, MAP & BAYESIAN INFERENCE

5.1 Bayesian Inference

5.2 Maximum Likelihood Estimation (MLE)

5.3 Maximum A-posteriori Method (MAP)

CHAPTER

SIX

AN ILLUSTRATIVE EXAMPLE FOR MLE AND MAP: LINEAR REGRESSION

6.1 Ordinary Least Squares

6.2 Ridge Regression

6.3 LASSO

CHAPTER

SEVEN

OPTIMIZATION METHODS

CHAPTER

EIGHT

MACHINE LEARNING WORKFLOW

Part III

Probabilistic Machine Learning

CHAPTER

NINE

MOTIVATION OF PROBABILISTIC MODELS

CHAPTER

TEN

KERNEL-BASED METHODS

10.1 Gaussian Processes

10.2 Additional Kernel-based Methods

CHAPTER

ELEVEN

OVERVIEW OF FURTHER PROBABILISTIC MODELS

Part IV

Applications

CHAPTER

TWELVE

BAYESIAN OPTIMIZATION

CHAPTER

THIRTEEN

QUANTIFICATION OF DESIGN UNCERTAINTY

CHAPTER

FOURTEEN

DATA-EFFICIENT REINFORCEMENT LEARNING