Chapter ML:IV

IV. Statistical Learning

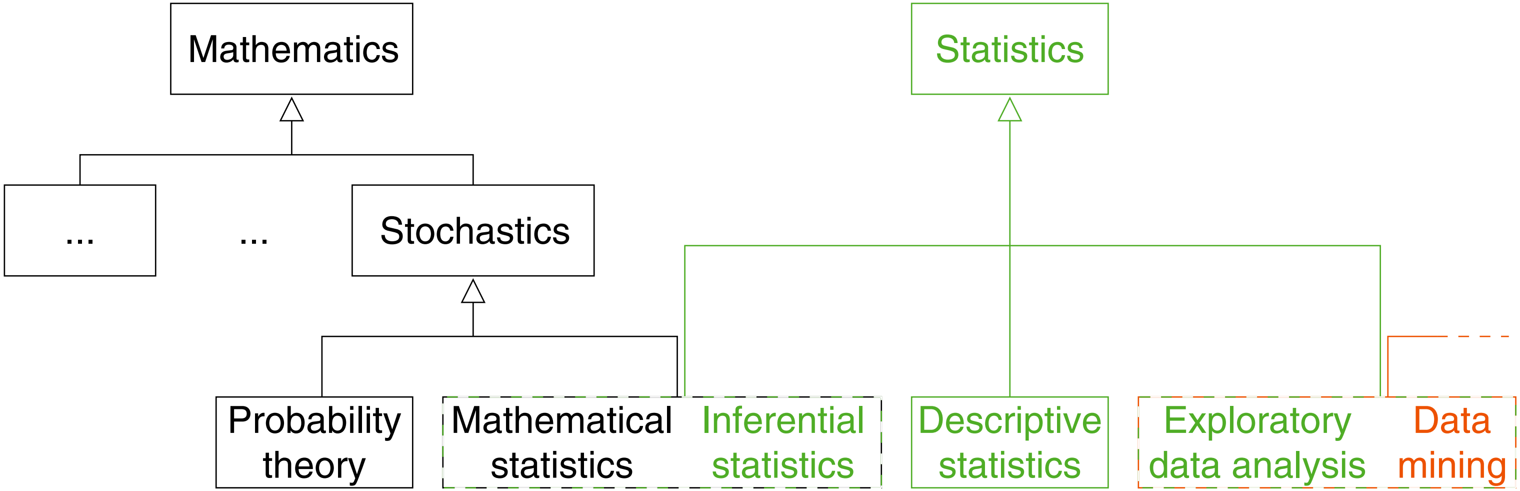
❑ Probability Basics

❑ Bayes Classification

❑ Maximum a-Posteriori Hypotheses

Probability Basics

Area Overview



From the area of probability theory:

- Kolmogorov Axioms

From the area of mathematical statistics:

- Naive Bayes

Probability Basics

Definition 1 (Random Experiment, Random Observation)

A random experiment or random trial is a procedure that, at least theoretically, can be repeated infinite times. It is characterized as follows:

1. Configuration.
   A precisely specified system that can be reconstructed.

2. Procedure.
   An instruction of how to execute the experiment, based on the configuration.

3. Unpredictability of the outcome.

Random experiments whose configuration and procedure are not designed artificially are called *natural random experiments* or *natural random observations*.

Remarks:

- ❑ A procedure can be repeated several times using the same system, but also with different "copies" of the original system.
- ❑ Random experiments are causal in the sense of cause and effect. The randomness of an experiment (the unpredictability of its outcome) is a consequence of the missing information about the causal chain. Hence a random experiment can turn into a deterministic process when new insights become known.

Probability Basics

Definition 2 (Sample Space, Event Space)

A set $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ is called sample space of a random experiment, if each experiment outcome is associated with at most one element $\omega \in \Omega$. The elements in Ω are called outcomes.

Let Ω be a finite sample space. Each subset $A \subseteq \Omega$ is called an event; an event A occurs iff the experiment outcome ω is a member of A. The set of all events, $\mathcal{P}(\Omega)$, is called the event space.

Probability Basics

Definition 3 (Important Event Types)

Let Ω be a finite sample space, and let $A \subseteq \Omega$ and $B \subseteq \Omega$ be two events. Then we agree on the following notation:

1. $\emptyset$ — The impossible event.
2. Ω — The certain event.
3. $\overline{A} := \Omega \setminus A$ — The complementary event (opposite event) of A.
4. $|A| = 1$ — An elementary event.

5. $A \subseteq B \quad \Leftrightarrow \quad A$ is a sub-event of B or "A entails B", $A \Rightarrow B$
6. $A = B \quad \Leftrightarrow \quad A \subseteq B$ and $B \subseteq A$
7. $A \cap B = \emptyset \quad \Leftrightarrow \quad A$ and B are incompatible (otherwise, they are compatible).

Probability Basics

Classical Concept Formation

Empirical law of large numbers:

Given a random experiment, the average of the outcomes obtained from a large number of trials is close to the expected value, and it will become closer as more trials are performed.

Probability Basics

Classical Concept Formation

Empirical law of large numbers:

Given a random experiment, the average of the outcomes obtained from a large number of trials is close to the expected value, and it will become closer as more trials are performed.

Definition 4 (Classical / Laplace Probability)

If each elementary event in Ω gets assigned the same probability (equiprobable events), then the probability $P(A)$ of an event A is defined as follows:

$$P(A) = \frac{|A|}{|\Omega|} = \frac{\text{number of cases favorable for } A}{\text{number of total outcomes possible}}$$

Remarks:

- A random experiment whose configuration and procedure imply an equiprobable sample space, be it by definition or by construction, is called Laplace experiment. The probabilities of the outcomes are called Laplace probabilities. Since they are defined by the experiment configuration along with the experiment procedure, they need not to be estimated.
- The assumption that a given experiment is a Laplace experiment is called Laplace assumption. If the Laplace assumption cannot be presumed, the probabilities can only be obtained from a (possibly large) number of trials.
- Strictly speaking, the Laplace probability as introduced above is not a definition but a circular definition: the probability concept is defined by means of the concept of equiprobability, i.e., another kind of probability.
- Inspired by the empirical law of large numbers, scientists have tried to develop a frequentist probability concept, which is based on the (fictitious) limit of the relative frequencies [von Mises, 1951]. These attempts failed since such a limit formation is possible only within mathematical settings (infinitesimal calculus), where accurate repetitions unto infinity can be made.

Probability Basics

Axiomatic Concept Formation

The basic steps of axiomatic concept formation:

(a) Postulate a function that assigns a probability to each element of the event space.

(b) Specify the required properties of this function in the form of axioms.

Probability Basics

Axiomatic Concept Formation

The basic steps of axiomatic concept formation:

(a) Postulate a function that assigns a probability to each element of the event space.

(b) Specify the required properties of this function in the form of axioms.

Definition 5 (Probability Measure [Kolmogorov 1933]**)**

Let Ω be a set, called sample space, and let $\mathcal{P}(\Omega)$ be the set of all events, called event space. A function P, $P : \mathcal{P}(\Omega) \rightarrow \mathbf{R}$, which maps each event $A \in \mathcal{P}(\Omega)$ onto a real number $P(A)$, is called probability measure if it has the following properties:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset$ implies $P(A \cup B) = P(A) + P(B)$ (Axiom III)

Probability Basics

Axiomatic Concept Formation

The basic steps of axiomatic concept formation:

(a) Postulate a function that assigns a probability to each element of the event space.

(b) Specify the required properties of this function in the form of axioms.

Definition 5 (Probability Measure [Kolmogorov 1933])

Let Ω be a set, called sample space, and let $\mathcal{P}(\Omega)$ be the set of all events, called event space. A function P, $P : \mathcal{P}(\Omega) \to \mathbf{R}$, which maps each event $A \in \mathcal{P}(\Omega)$ onto a real number $P(A)$, is called probability measure if it has the following properties:

1. $P(A) \geq 0$ (Axiom I)
2. $P(\Omega) = 1$ (Axiom II)
3. $A \cap B = \emptyset \quad \rightarrow \quad P(A \cup B) = P(A) + P(B)$ (Axiom III)

Probability Basics

Axiomatic Concept Formation (continued)

Definition 6 (Probability Space)

Let Ω be a sample space, let $\mathcal{P}(\Omega)$ be an event space, and let $P : \mathcal{P}(\Omega) \rightarrow \mathbf{R}$ be a probability measure. Then the tuple (Ω, P), as well as the triple $(\Omega, \mathcal{P}(\Omega), P)$, is called probability space.

Probability Basics

Axiomatic Concept Formation (continued)

Definition 6 (Probability Space)

Let Ω be a sample space, let $\mathcal{P}(\Omega)$ be an event space, and let $P : \mathcal{P}(\Omega) \rightarrow \mathbf{R}$ be a probability measure. Then the tuple (Ω, P), as well as the triple $(\Omega, \mathcal{P}(\Omega), P)$, is called probability space.

Theorem 7 (Implications of Kolmogorov Axioms)

1. $P(A) + P(\overline{A}) = 1$ (from Axioms II, III)

2. $P(\emptyset) = 0$ (from 1. with $A = \Omega$)

3. Monotonicity law of the probability measure:
   $A \subseteq B \;\Rightarrow\; P(A) \leq P(B)$ (from Axioms I, II)

4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (from Axiom III)

5. Let $A_1, A_2 \ldots, A_k$ be mutually exclusive (incompatible), then holds:
   $P(A_1 \cup A_2 \cup \ldots \cup A_k) = P(A_1) + P(A_2) + \ldots + P(A_k)$

Remarks:

- The three axioms are also called the axiom system of Kolmogorov.
- $P(A)$ is denoted as the “probability of the occurrence of A”
- Observe that nothing is said about how to interpret the probabilities P.
- Also observe that nothing is said about the distribution of the probabilities P.
- A function that provides the three properties of a probability measure is called a non-negative, normalized, additive measure.

Probability Basics

Conditional Probability

Definition 8 (Conditional Probability)

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event A given that event B is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called probability of A under condition B.

Probability Basics

Conditional Probability

Definition 8 (Conditional Probability)

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event A given that event B is known to have occurred is defined as follows:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

$P(A \mid B)$ is called probability of A under condition B.

Probability Basics

Conditional Probability (continued)

Theorem 9 (Total Probability)

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for an $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$

Probability Basics

Conditional Probability (continued)

Theorem 9 (Total Probability)

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k$ be mutually exclusive events with $\Omega = A_1 \cup \ldots \cup A_k$, $P(A_i) > 0$, $i = 1, \ldots, k$. Then for an $B \in \mathcal{P}(\Omega)$ holds:

$$P(B) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)$$

Proof

$$\begin{aligned}
P(B) &= P(\Omega \cap B) \\
&= P((A_1 \cup \ldots \cup A_k) \cap B) \\
&= P((A_1 \cap B) \cup \ldots \cup (A_k \cap B)) \\
&= \sum_{i=1}^{k} P(A_i \cap B) \\
&= \sum_{i=1}^{k} P(B \cap A_i) \;=\; \sum_{i=1}^{k} P(A_i) \cdot \underline{P(B \mid A_i)}
\end{aligned}$$

Remarks:

- The theorem of total probability states that the probability of an arbitray event equals the sum of the probabilities of the sub-events into which the event has been partitioned.
- Considered as a function in the parameter A and the constant B, the conditional probability $P(A \mid B)$ fulfills the Kolmogorov axioms and in turn defines a probability measure, denoted as P_B here.
- Important consequences (deductions) from the conditional probability definition:

  1. $P(A \cap B) = P(B) \cdot P(A \mid B)$ (see multiplication rule in Definition 10)
  2. $P(A \cap B) = P(B \cap A) = P(A) \cdot P(B \mid A)$
  3. $P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \Leftrightarrow P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{P(A) \cdot P(B \mid A)}{P(B)}$
  4. $P(\overline{A} \mid B) = 1 - P(A \mid B)$ or: $P_B(\overline{A}) = 1 - P_B(A)$ (see Point 1 in Theorem 7)

- While the consequence 4 is obvious since P_B is a probability measure, the interpretation of complementary events when used as conditions may be confusing. In particular, the following inequality must be assumed: $P(A \mid \overline{B}) \neq 1 - P(A \mid B)$

  For illustrating purposes, let $P(A \mid B) = 0.9$ for the event “Road is wet” (A) under the event “It’s raining” (B). Then this information doesn’t give us any knowledge regarding the wetness of the road under the complementary event $\overline{B}$ “It’s not raining”.

Probability Basics

Independence of Events

Definition 10 (Statistical Independence of two Events)

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then A and B are called statistically independent iff the following equation holds:

$$P(A \cap B) = P(A) \cdot P(B) \qquad \text{“multiplication rule”}$$

If statistical independence is given for A, B, and $0 < P(B) < 1$, the following equivalences hold:

$$\begin{aligned} & P(A \cap B) &= P(A) \cdot P(B) \\ \Leftrightarrow \quad & P(A \mid B) &= P(A \mid \overline{B}) \\ \Leftrightarrow \quad & P(A \mid B) &= P(A) \end{aligned}$$

Probability Basics

Independence of Events (continued)

Definition 11 (Statistical Independence of k Events)

Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A_1, \ldots, A_k \in \mathcal{P}(\Omega)$ be events. Then the $A_1, \ldots, A_k$ are called jointly statistically independent at P iff for all subsets $\{A_{i_1}, \ldots, A_{i_l}\} \subseteq \{A_1, \ldots, A_k\}$ the multiplication rule holds:

$$P(A_{i_1} \cap \ldots \cap A_{i_l}) = P(A_{i_1}) \cdot \ \ldots \ \cdot P(A_{i_l}),$$

where $i_1 < i_2 < \ldots < i_l$ and $2 \leq l \leq k$.