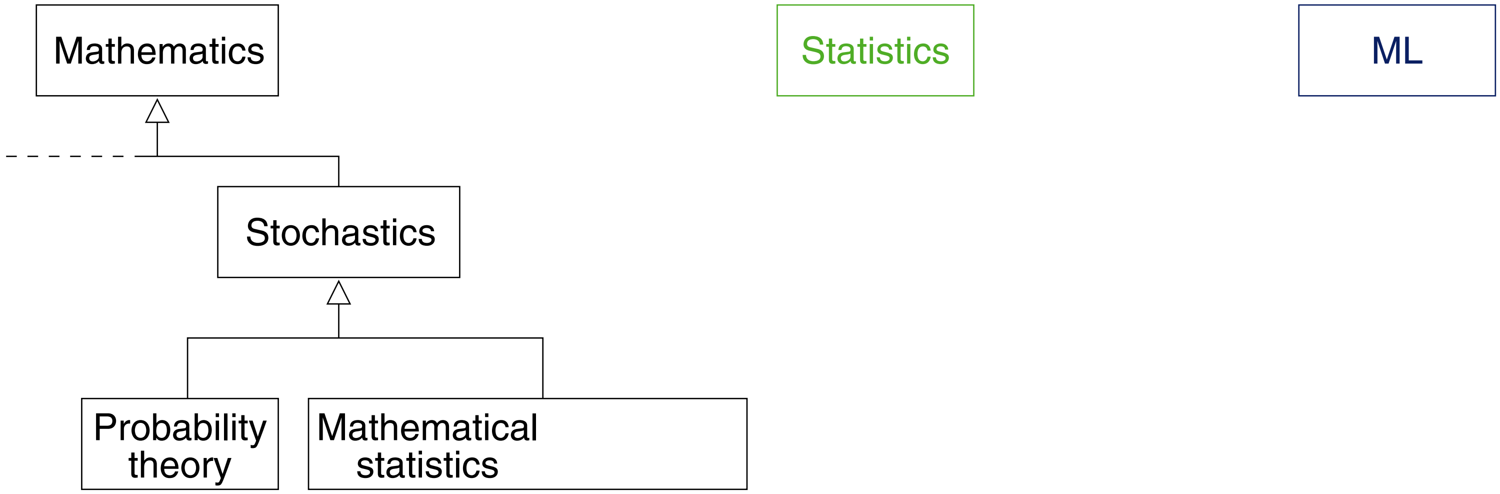
Chapter ML:IV

IV. Statistical Learning

- Probability Basics
- Bayes Classification
- MAP versus ML Hypotheses

Probability Basics

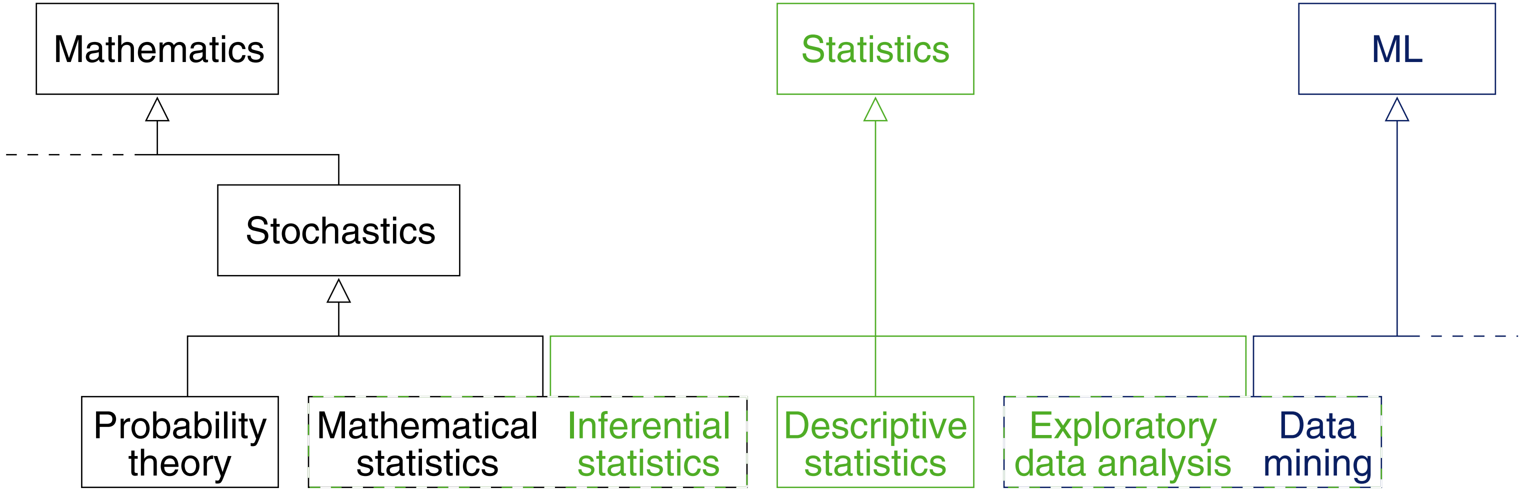
Area Overview



- ❑ Probability theory: probability measures, Kolmogorov axioms
- ❑ Mathematical statistics: application of probability theory, Naive Bayes

Probability Basics

Area Overview



- Probability theory: probability measures, Kolmogorov axioms
- Mathematical statistics: application of probability theory, Naive Bayes

- Inferential statistics: hypothesis tests, confidence intervals
- Descriptive statistics: variances, contingencies
- Exploratory data analysis: histograms, principal component analysis

- Data mining: anomaly detection, cluster analysis

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.

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.

  In particular, a random experiment is called *ergodic* if its time average (= sequential analysis) is the same as its ensemble average (= parallel analysis). [Wikipedia]

- Note that random experiments are causal in the sense of cause and effect. The randomness of an experiment, i.e., 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 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.

Examples:

Experiment:	Rolling a dice.
Sample space:	$\Omega = \{1, 2, 3, 4, 5, 6\}$
Some event:	$A = \{2, 4, 6\}$

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.

Examples:

Experiment:	Rolling a dice.	Rolling two dice at the same time.
Sample space:	$\Omega = \{1, 2, 3, 4, 5, 6\}$	$\Omega = \{\{1,1\}, \{1,2\}, \ldots, \{2,2\}, \ldots, \{6,6\}\}$
Some event:	$A = \{2, 4, 6\}$	$B = \{\{1, 2\}\}$

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.

Examples:

Experiment:	Rolling a dice.	Rolling two dice at the same time.
Sample space:	$\Omega = \{1, 2, 3, 4, 5, 6\}$	$\Omega = \{\{1,1\}, \{1,2\}, \ldots, \{2,2\}, \ldots, \{6,6\}\}$
Some event:	$A = \{2, 4, 6\}$	$B = \{\{1,2\}\}$

Rolling two dice in succession.

$\Omega = \{(1,1), (1,2), \ldots, (2,1), \ldots, (6,6)\}$

$B = \{(1,2), (2,1)\}$

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 of A.
4. $|A| = 1$ An elementary event.
5. $A \subseteq B$ $\Leftrightarrow$ A is a sub-event of B, "A entails B", $A \Rightarrow B$
6. $A = B$ $\Leftrightarrow$ $A \subseteq B$ and $B \subseteq A$
7. $A \cap B = \emptyset$ $\Leftrightarrow$ A and B are incompatible (otherwise, they are compatible).

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 of A.
4. $|A| = 1$ An elementary event.
5. $A \subseteq B$ $\Leftrightarrow$ A is a sub-event of B, "A entails B", $A \Rightarrow B$
6. $A = B$ $\Leftrightarrow$ $A \subseteq B$ and $B \subseteq A$
7. $A \cap B = \emptyset$ $\Leftrightarrow$ A and B are incompatible (otherwise, they are compatible).

Example (Point 5) :

"Roll a two." $\subset$ "Even number roll."

"2" entails "Even number roll."

"2" $\Rightarrow$ "2 or 4 or 6"

Remarks:

- Alternative and semantically equivalent notations of the probability for the combined event “A and B” :

  1. $P(A, B)$
  2. $P(A \wedge B)$
  3. $P(A \cap B)$

Probability Basics

Approaches to Capture the Nature of Probability

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Probability Basics

Approaches to Capture the Nature of Probability

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Definition 4 (Classical / Laplace Probability [1749-1827])

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 Laplace probabilities 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.

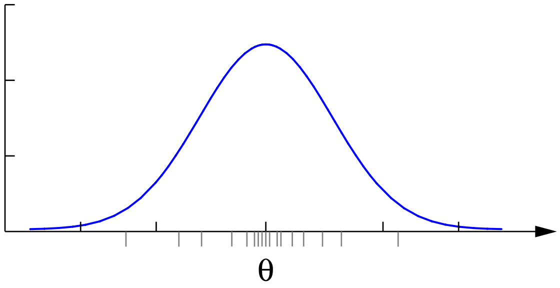
Probability Basics

Approaches to Capture the Nature of Probability (continued)

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Basis is the 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.



Remarks:

- Inspired by the empirical law of large numbers, scientists have tried to develop a frequentist probability concept, which is completely 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

Approaches to Capture the Nature of Probability (continued)

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Integration of previous knowledge into a decision process:

$$p(\text{hypothesis} \mid \text{data}) = \frac{p(\text{data} \mid \text{hypothesis}) \cdot p(\text{hypothesis})}{p(\text{data})}$$

Probability Basics

Approaches to Capture the Nature of Probability (continued)

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Integration of previous knowledge into a decision process:

$$p(\text{hypothesis} \mid \text{data}) = \frac{p(\text{data} \mid \text{hypothesis}) \cdot p(\text{hypothesis})}{p(\text{data})}$$

- Likelihood: How accurate does a hypothesis “explain” (fit) the data?
- Prior: How probable is a hypothesis a-priori?

Probability Basics

Approaches to Capture the Nature of Probability (continued)

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Integration of previous knowledge into a decision process:

$$p(\text{hypothesis} \mid \text{data}) = \frac{p(\text{data} \mid \text{hypothesis}) \cdot p(\text{hypothesis})}{p(\text{data})}$$

- ❑ Likelihood: How accurate does a hypothesis “explain” (fit) the data?
- ❑ Prior: How probable is a hypothesis a-priori?

$$p(\text{hypothesis} \mid \text{data}) \propto p(\text{data} \mid \text{hypothesis}) \cdot p(\text{hypothesis})$$

Remarks:

- Likelihood is the hypothetical probability that an event that has already occurred would yield a specific outcome. The concept differs from that of a probability in that a probability refers to the occurrence of future events, while a likelihood refers to past events with known outcomes. [Mathworld]

- If applicable and if properly applied the frequentist and the Bayesian approach lead to the same result in most cases.

- The frequentism approach cannot handle singleton or rare events. Example: "What are the chances that the first human mission to Mars will become a success?"

- "It is unanimously agreed that statistics depends somehow on probability. But, as to what probability is and how it is connected with statistics, there has seldom been such complete disagreement and breakdown of communication since the Tower of Babel. Doubtless, much of the disagreement is merely terminological and would disappear under sufficiently sharp analysis." [Savage, 1954]

Probability Basics

Approaches to Capture the Nature of Probability (continued)

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Axiomatic approach to phenomema modeling:

(a) Postulate a function that assigns a “probability” to each element in $\mathcal{P}(\Omega)$.

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

Probability Basics

Approaches to Capture the Nature of Probability (continued)

1. Classical definition, symmetry principle
2. Frequentism
3. Subjectivism, Bayesian probability
4. Axiomatic probability

Axiomatic approach to phenomema modeling:

(a) Postulate a function that assigns a “probability” to each element in $\mathcal{P}(\Omega)$.

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

Probability Basics

Axiomatic Approach to Probability

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 Approach to Probability

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 Approach to Probability (continued)

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. $A \cap B = \emptyset \quad \to \quad P(A \cup B) = P(A) + P(B)$ (Axiom III)

Definition 6 (Probability Space)

Let Ω be a sample space, let $\mathcal{P}(\Omega)$ be an event space, and let $P : \mathcal{P}(\Omega) \to \mathbf{R}$ 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 Approach to Probability (continued)

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. $A \cap B = \emptyset \quad \rightarrow \quad P(A \cup B) = P(A) + P(B)$ (Axiom III)

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.

We can work with probabilities without interpreting them.

Probability Basics

Axiomatic Approach to Probability (continued)

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. “Sum rule” or “addition rule” :
   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (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 called “probability of the occurrence of A.”
- Observe that nothing is said about how to interpret the probabilities P. An axiomatic approach does not explain but “only” specifies properties.
- 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, and 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 (continued)

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)

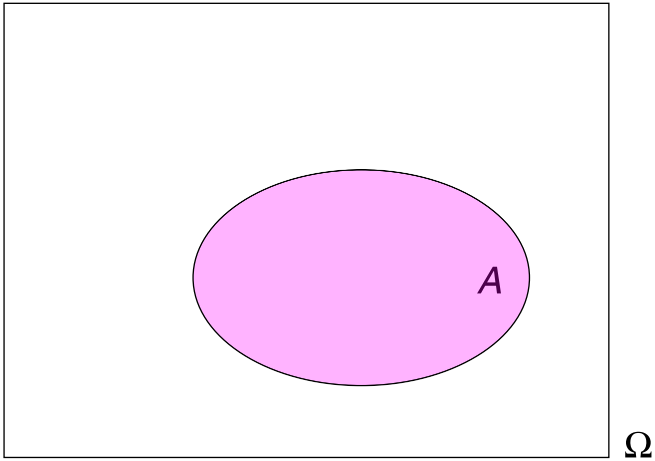
Definition 8 (Conditional Probability)

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."

A : The road is wet.



$A \equiv A \mid \Omega$

Probability Basics

Conditional Probability (continued)

Definition 8 (Conditional Probability)

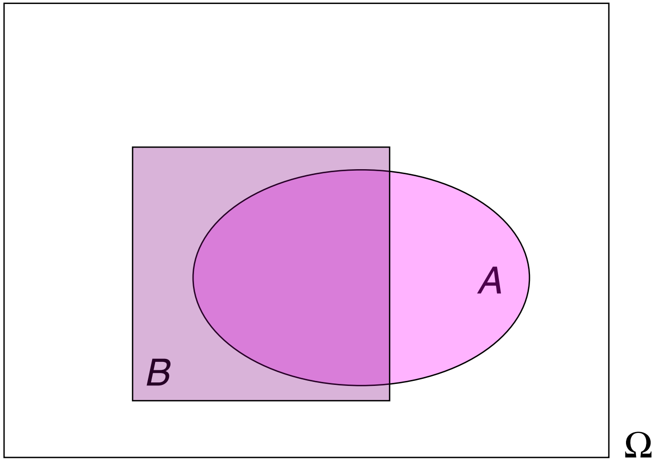
Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then the probability of the occurrence of event A given that event B is known to have occurred is defined as follows:

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

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

A : The road is wet.

B : It's raining.



$B \equiv B \mid \Omega$

Probability Basics

Conditional Probability (continued)

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:

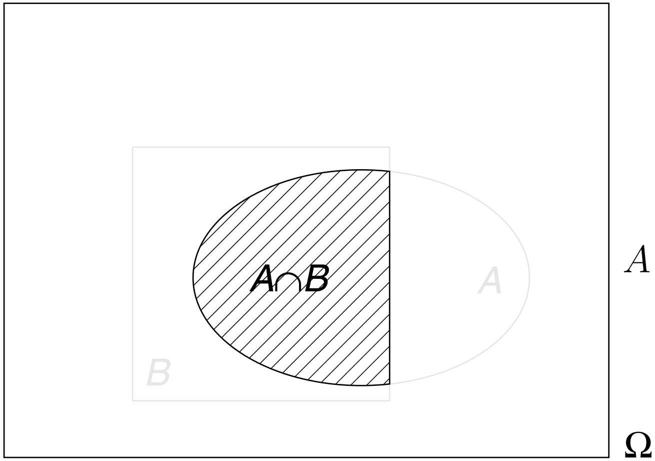
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0$$

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

A : The road is wet.

B : It’s raining.

$A \cap B$: The road is wet and it’s raining.



$A \cap B \equiv A \cap B \mid \Omega$

Probability Basics

Conditional Probability (continued)

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:

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

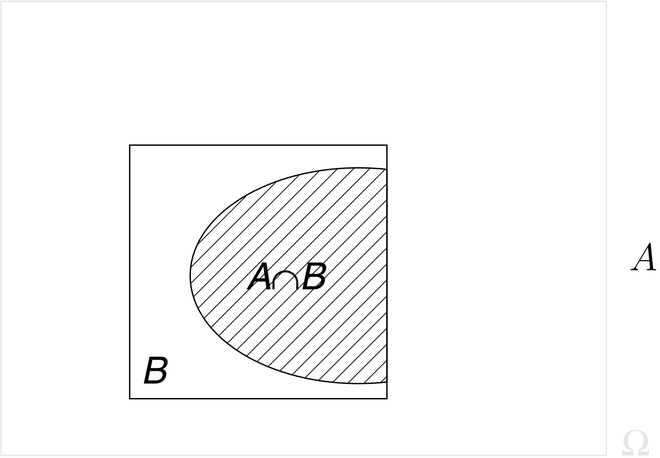
$P(A \mid B)$ is called "probability of A under condition B."

A : The road is wet.

B : It's raining.

$A \cap B$: The road is wet and it's raining.

$A \mid B$: The road is wet when it's raining.



$A \mid B \equiv A \cap B \mid B$

Remarks:

- Important consequences (deductions) from the conditional probability definition:

  1. $P(A \cap B) = P(B) \cdot P(A \mid B)$ (see multiplication rule for statistical independence)

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

  3. $P(B) \cdot P(A \mid B) = P(A) \cdot P(B \mid A) \Leftrightarrow P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} \overset{(\star)}{=} \dfrac{P(A) \cdot P(B \mid A)}{P(B)}$

  4. $P(\overline{A} \mid B) = 1 - P(A \mid B)$ (see Point 1 in Kolmogorov implications)

     or $P_B(\overline{A}) = 1 - P_B(A)$

$(\star)$ The identity shows the (simple) Bayes rule.

- 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.

Remarks (continued) :

- While Deduction 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, consider the probability $P(A \mid B) = 0.9$ for the event “The road is wet” (A) under the event “It’s raining” (B). Observe that 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

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 each $B \in \mathcal{P}(\Omega)$ holds:

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

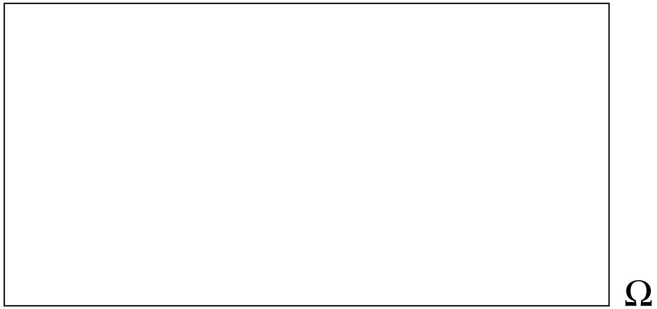
Probability Basics

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 each $B \in \mathcal{P}(\Omega)$ holds:

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



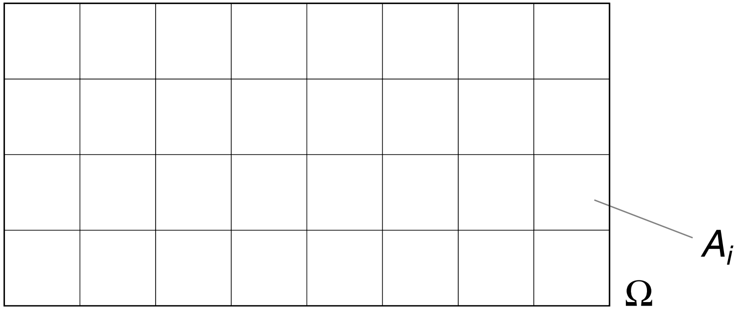
Probability Basics

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 each $B \in \mathcal{P}(\Omega)$ holds:

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



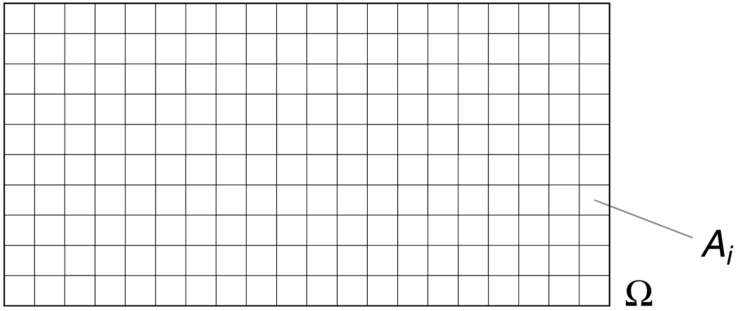
Probability Basics

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 each $B \in \mathcal{P}(\Omega)$ holds:

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



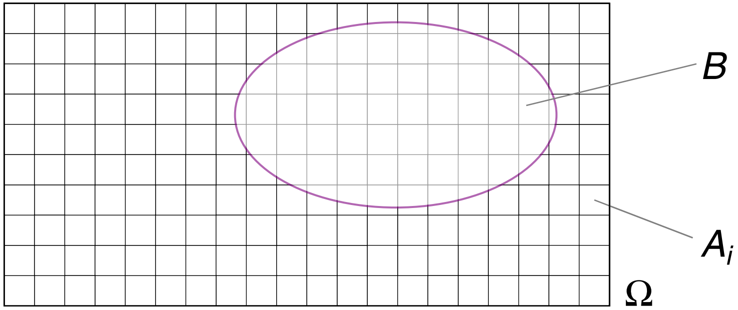
Probability Basics

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 each $B \in \mathcal{P}(\Omega)$ holds:

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



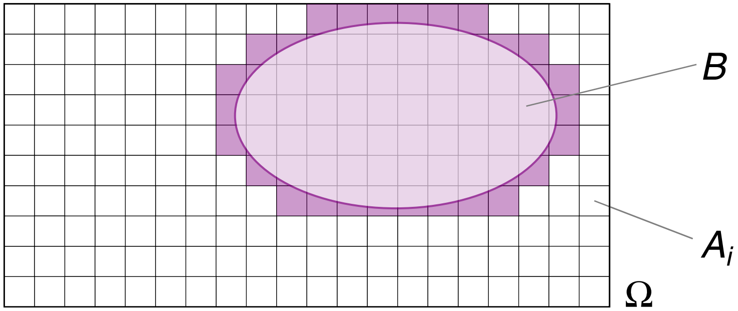
Probability Basics

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 each $B \in \mathcal{P}(\Omega)$ holds:

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



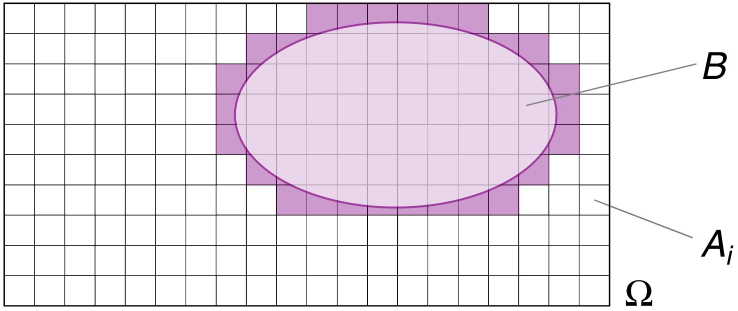
Probability Basics

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 each $B \in \mathcal{P}(\Omega)$ holds:

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



Proof

$$\begin{aligned}
P(B) &= P(\Omega \cap B) \\
&= P((A_1 \cup \ldots \cup A_k) \cap B) && \text{(exploitation of completeness of the } A_i\text{)} \\
&= P((A_1 \cap B) \cup \ldots \cup (A_k \cap B)) && \text{(exploitation of exclusiveness of the } A_i\text{)} \\
&= \sum_{i=1}^{k} P(A_i \cap B) = \sum_{i=1}^{k} P(B \cap A_i) = \sum_{i=1}^{k} P(A_i) \cdot P(B \mid A_i)
\end{aligned}$$

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.

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:

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

Probability Basics

Independence of Events (continued)

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:

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

Probability Basics

Independence of Events (continued)

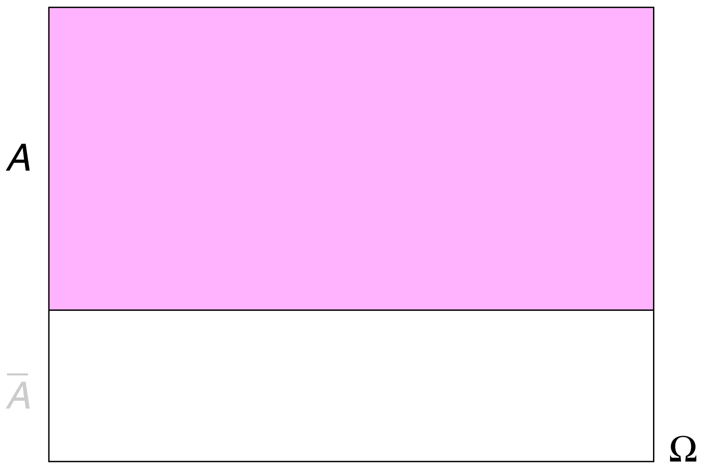
Definition 10 (Statistical Independence of two Events)

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:

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



Probability Basics

Independence of Events (continued)

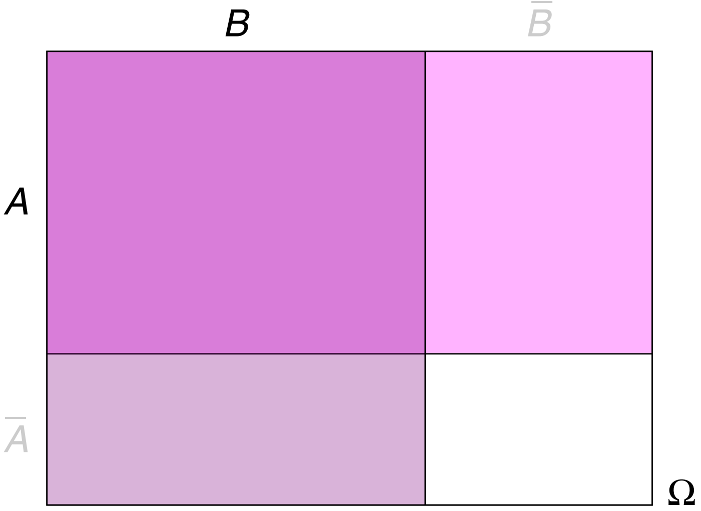
Definition 10 (Statistical Independence of two Events)

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:

$$P(A \cap B) = P(A) \cdot P(B)$$ “multiplication rule”

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

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



[dependent events]

Probability Basics

Independence of Events (continued)

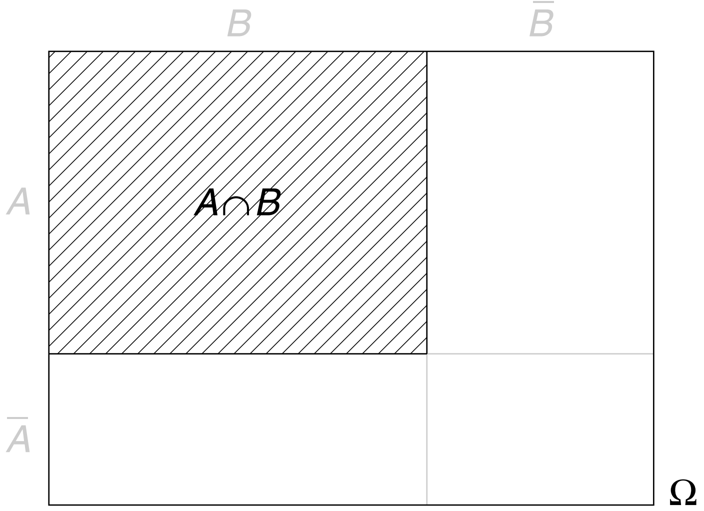
Definition 10 (Statistical Independence of two Events)

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:

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



[dependent events]

Probability Basics

Independence of Events (continued)

Definition 10 (Statistical Independence of two Events)

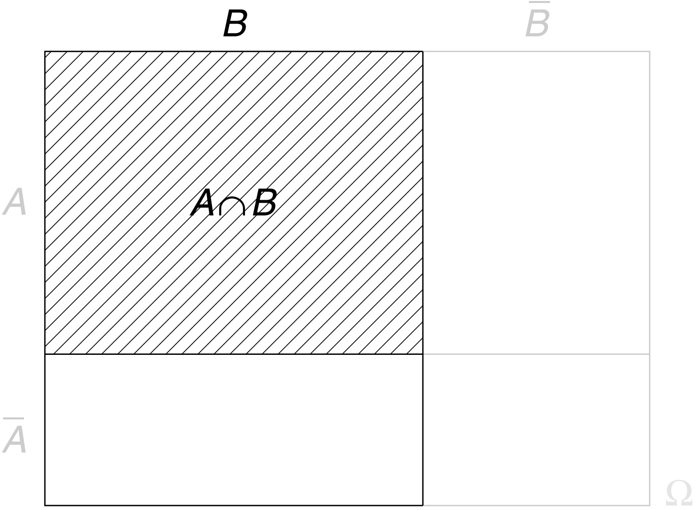
Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then A and B are called statistically independent iff the following equation holds:

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

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

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

[dependent events]



Probability Basics

Independence of Events (continued)

Definition 10 (Statistical Independence of two Events)

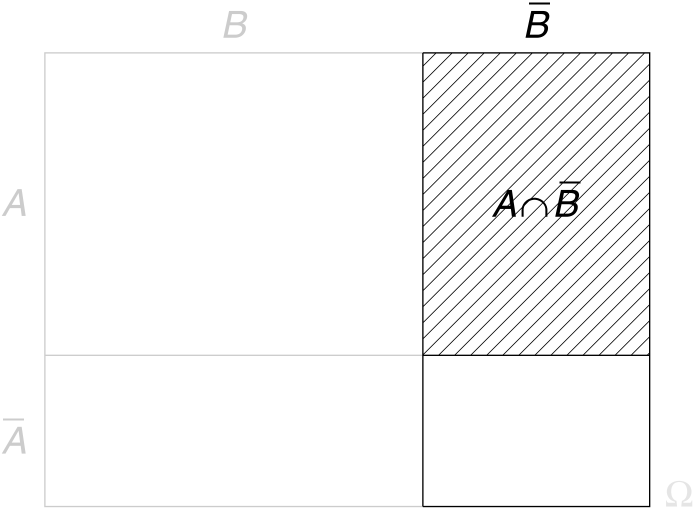
Let $(\Omega, \mathcal{P}(\Omega), P)$ be a probability space, and let $A, B \in \mathcal{P}(\Omega)$ be two events. Then A and B are called statistically independent iff the following equation holds:

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

"multiplication rule"

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

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



[dependent events]

Probability Basics

Independence of Events (continued)

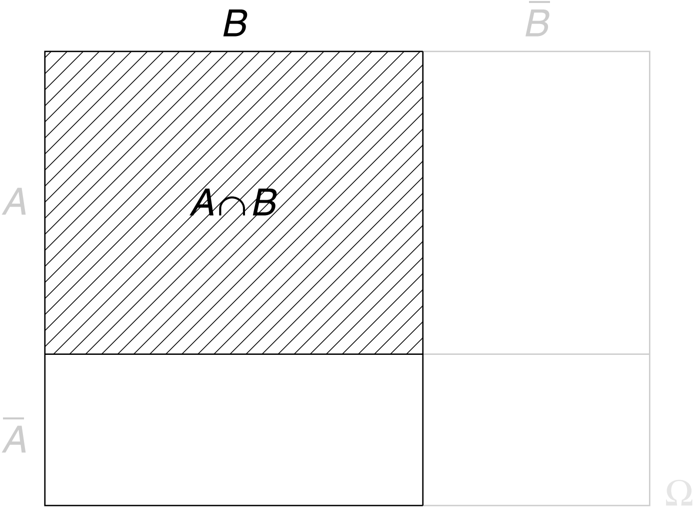
Definition 10 (Statistical Independence of two Events)

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:

$$P(A \cap B) = P(A) \cdot P(B)$$ “multiplication rule”

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

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



[dependent events]

Probability Basics

Independence of Events (continued)

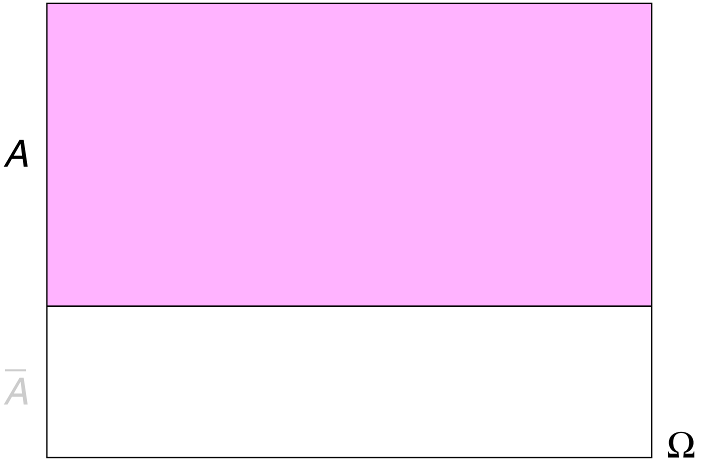
Definition 10 (Statistical Independence of two Events)

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:

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



[dependent events]

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 k 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 $2 \leq l \leq k$.

Chapter ML:IV (continued)

IV. Statistical Learning

- ❑ Probability Basics
- ❑ Bayes Classification
- ❑ MAP versus ML Hypotheses

Bayes Classification

Generative Approach to Classification Problems

Setting:

- X is a set of feature vectors.
- C is a set of classes.
- $c : X \to C$ is the (unknown) ideal classifier for X.
- $D = \{(\mathbf{x}_1, c(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, c(\mathbf{x}_n))\} \subseteq X \times C$ is a set of examples.

Todo:

- Approximate $c(\mathbf{x})$, which is implicitly given via D, by estimating the underlying joint probability $P(\mathbf{x}, c(\mathbf{x}))$.

Bayes Classification

Bayes Theorem

Theorem 12 (Bayes [1701-1761]**)**

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 event $B \in \mathcal{P}(\Omega)$ with $P(B) > 0$ holds:

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

$P(A_i)$ is called *prior probability* of A_i.

$P(A_i \mid B)$ is called *posterior probability* of A_i.

Bayes Classification

Bayes Theorem (continued)

Proof (Bayes Theorem)

From the conditional probabilities for $P(B \mid A_i)$ and $P(A_i \mid B)$ follows:

$$P(A_i \mid B) = \frac{P(B \cap A_i)}{P(B)} = \frac{P(A_i) \cdot P(B \mid A_i)}{P(B)}$$

Bayes Classification

Bayes Theorem (continued)

Proof (Bayes Theorem)

From the conditional probabilities for $P(B \mid A_i)$ and $P(A_i \mid B)$ follows:

$$P(A_i \mid B) \;=\; \frac{P(B \cap A_i)}{P(B)} \;=\; \frac{P(A_i) \cdot P(B \mid A_i)}{P(B)}$$

Applying the theorem of total probability to $P(B)$,

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

will yield the claim.

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009]

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

   B : *test_pos*

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

$\Rightarrow$ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

B : *test_pos*

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

⇒ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

B : *test_pos*

2. $B \mid A_1$: *test_pos* | *Aids* $P(B \mid A_1) = 0.98$ (result from clinical trials)

3. $B \mid A_2$: *test_pos* | *no_Aids* $P(B \mid A_2) = 0.03$ (result from clinical trials)

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

   ⇒ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

   B : *test_pos*

2. $B \mid A_1$: *test_pos* | *Aids* $P(B \mid A_1) = 0.98$ (result from clinical trials)

3. $B \mid A_2$: *test_pos* | *no_Aids* $P(B \mid A_2) = 0.03$ (result from clinical trials)

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

$\Rightarrow$ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

B : *test_pos* $\Rightarrow$ $P(B) = \sum_{i=1}^{2} P(A_i) \cdot P(B \mid A_i) = 0.031$

2. $B \mid A_1$: *test_pos* | *Aids* $P(B \mid A_1) = 0.98$ (result from clinical trials)

3. $B \mid A_2$: *test_pos* | *no_Aids* $P(B \mid A_2) = 0.03$ (result from clinical trials)

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

$\Rightarrow$ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

B : *test_pos* $\Rightarrow$ $P(B) = \sum_{i=1}^{2} P(A_i) \cdot P(B \mid A_i) = 0.031$

2. $B \mid A_1$: *test_pos* | *Aids* $P(B \mid A_1) = 0.98$ (result from clinical trials)

3. $B \mid A_2$: *test_pos* | *no_Aids* $P(B \mid A_2) = 0.03$ (result from clinical trials)

Simple Bayes formula:

$$P(\textit{Aids} \mid \textit{test\_pos}) = P(A_1 \mid B) = \frac{P(A_1) \cdot P(B \mid A_1)}{P(B)}$$

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

⇒ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

B : *test_pos* ⇒ $P(B) = \sum_{i=1}^{2} P(A_i) \cdot P(B \mid A_i) = 0.031$

2. $B \mid A_1$: *test_pos* | *Aids* $P(B \mid A_1) = 0.98$ (result from clinical trials)

3. $B \mid A_2$: *test_pos* | *no_Aids* $P(B \mid A_2) = 0.03$ (result from clinical trials)

Simple Bayes formula:

$$P(\textit{Aids} \mid \textit{test\_pos}) = P(A_1 \mid B) = \frac{P(A_1) \cdot P(B \mid A_1)}{P(B)} = \frac{0.001 \cdot 0.98}{0.031} = 0.032 = 3.2\%$$

Bayes Classification

Example: Reasoning About a Disease [Kirchgessner 2009] (continued)

1. A_1 : *Aids* $P(A_1) = 0.001$ (prior knowledge about population)

⇒ A_2 : *no_Aids* $P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

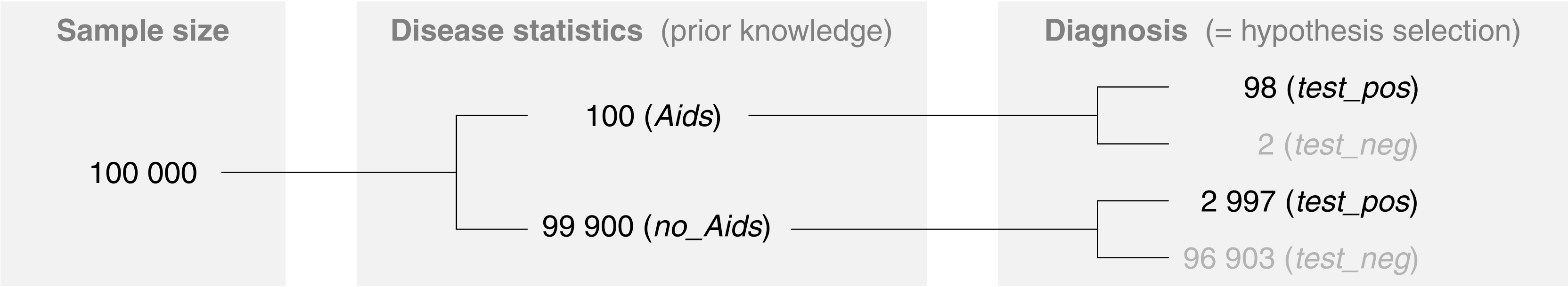
B : *test_pos* ⇒ $P(B) = \sum_{i=1}^{2} P(A_i) \cdot P(B \mid A_i) = 0.031$

2. $B \mid A_1$: *test_pos* | *Aids* $P(B \mid A_1) = 0.98$ (result from clinical trials)

3. $B \mid A_2$: *test_pos* | *no_Aids* $P(B \mid A_2) = 0.03$ (result from clinical trials)

Simple Bayes formula:

$$P(\textit{Aids} \mid \textit{test\_pos}) = P(A_1 \mid B) = \frac{P(A_1) \cdot P(B \mid A_1)}{P(B)} = \frac{0.001 \cdot 0.98}{0.031} = 0.032 = 3.2\%$$



Sample size

100 000

Disease statistics (prior knowledge)

100 (*Aids*)

99 900 (*no_Aids*)

Diagnosis (= hypothesis selection)

98 (*test_pos*)

2 (*test_neg*)

2 997 (*test_pos*)

96 903 (*test_neg*)

Bayes Classification

Combined Conditional Events: $P(A_i \mid B_1, \ldots, B_p)$

Let $P(A_i \mid B_1, \ldots, B_p)$ denote the probability of the occurrence of event A_i given that the events (conditions) $B_1, \ldots, B_p$ are known to have occurred.

Bayes Classification

Combined Conditional Events: $P(A_i \mid B_1, \ldots, B_p)$

Let $P(A_i \mid B_1, \ldots, B_p)$ denote the probability of the occurrence of event A_i given that the events (conditions) $B_1, \ldots, B_p$ are known to have occurred.

Applied to a classification problem:

- A_i corresponds to an event of the kind "*class=c$_i$*", the B_j, $j = 1, \ldots, p$, correspond to p events of the kind "*attribute$_j$=x$_j$*".
- observable connection (in the regular situation) : $B_1, \ldots, B_p \mid A_i$
- reversed connection (in a diagnosis situation) : $A_i \mid B_1, \ldots, B_p$

Bayes Classification

Combined Conditional Events: $P(A_i \mid B_1, \ldots, B_p)$

Let $P(A_i \mid B_1, \ldots, B_p)$ denote the probability of the occurrence of event A_i given that the events (conditions) $B_1, \ldots, B_p$ are known to have occurred.

Applied to a classification problem:

- A_i corresponds to an event of the kind "*class=c*$_i$",
  the B_j, $j = 1, \ldots, p$, correspond to p events of the kind "*attribute*$_j$*=x*$_j$".
- observable connection (in the regular situation) : $B_1, \ldots, B_p \mid A_i$
- reversed connection (in a diagnosis situation) : $A_i \mid B_1, \ldots, B_p$

If sufficient data for estimating $P(A_i)$ and $P(B_1, \ldots, B_p \mid A_i)$ is provided, then $P(A_i \mid B_1, \ldots, B_p)$ can be computed with the Theorem of Bayes:

$$P(A_i \mid B_1, \ldots, B_p) = \frac{P(A_i) \cdot P(B_1, \ldots, B_p \mid A_i)}{P(B_1, \ldots, B_p)} \qquad (\star)$$

Remarks [information gain for classification] :

- How probability theory is applied to classification problem solving:
  - Classes and attribute-value pairs are interpreted as events. The relation to an underlying sample space Ω, $\Omega = \{\omega_1, \ldots, \omega_n\}$, from which the events are subsets, is not considered.
  - Observable or measurable and possibly causal connection: It is (or was in the past) regularly observed that in situation A_i (e.g. a disease) the symptoms $B_1, \ldots, B_p$ occur. One may denote this as forward connection.
  - Reversed connection, typically an analysis or diagnosis situation: The symptoms $B_1, \ldots, B_p$ are observed, and one is interested in the probability that A_i is given or has occurred.
  - Based on the prior probabilities of the classes (aka class priors), $P(\mathit{class}{=}c_i)$, and the class-conditional probabilities of the observable connections (aka likelihoods), $P(\mathit{attribute}_1{=}x_1, \ldots, \mathit{attribute}_p{=}x_p \mid \mathit{class}{=}c_i)$, the conditional class probabilities in an analysis situation, $P(\mathit{class}{=}c_i \mid \mathit{attribute}_1{=}x_1, \ldots, \mathit{attribute}_p{=}x_p)$, can be computed with the Theorem of Bayes.

- Note that a class-conditional event “$\mathit{attribute}_j{=}\mathrm{x}_j \mid \mathit{class}{=}c_i$” does not necessarily model a cause-effect relation: the event “$\mathit{class}{=}c_i$” may cause—but does not need to cause—the event “$\mathit{attribute}_j{=}x_j$”.

Remarks (continued) :

- Recap. Alternative and semantically equivalent notations of $P(A_i \mid B_1, \ldots, B_p)$:

  1. $P(A_i \mid B_1, \ldots, B_p)$
  2. $P(A_i \mid B_1 \wedge \ldots \wedge B_p)$
  3. $P(A_i \mid B_1 \cap \ldots \cap B_p)$

Bayes Classification

Naive Bayes

The compilation of a database from which reliable values for the $P(B_1, \ldots, B_p \mid A_i)$ can be obtained is often infeasible. The way out:

(a) Naive Bayes Assumption: “Given condition A_i, the $B_1, \ldots, B_p$ are statistically independent” (aka the B_j are *conditionally independent*). Notation:

$$P(B_1, \ldots, B_p \mid A_i) \overset{NB}{=} \prod_{j=1}^{p} P(B_j \mid A_i)$$

Bayes Classification

Naive Bayes

The compilation of a database from which reliable values for the $P(B_1, \ldots, B_p \mid A_i)$ can be obtained is often infeasible. The way out:

(a) Naive Bayes Assumption: “Given condition A_i, the $B_1, \ldots, B_p$ are statistically independent” (aka the B_j are *conditionally independent*). Notation:

$$P(B_1, \ldots, B_p \mid A_i) \overset{NB}{=} \prod_{j=1}^{p} P(B_j \mid A_i)$$

(b) Given a set $\{A_1, \ldots, A_k\}$ of alternative events (causes or classes), the most probable event under the Naive Bayes assumption, A_{NB}, can be computed with the Theorem of Bayes $(\star)$:

$$\underset{A_i \in \{A_1,\ldots,A_k\}}{\text{argmax}} \frac{P(A_i) \cdot P(B_1, \ldots, B_p \mid A_i)}{P(B_1, \ldots, B_p)} \overset{NB}{=} \underset{A_i \in \{A_1,\ldots,A_k\}}{\text{argmax}} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i) = A_{NB}$$

Remarks:

- Rationale for the Naive Bayes Assumption. Usually the probability $P(B_1, \ldots, B_p \mid A_i)$ cannot be estimated: Suppose that we are given p attributes (features) and that the domains of the attributes contain minimum l values each.

  Then, for as many as l^p different feature vectors the probabilities $P(B_{1=x_1}, \ldots, B_{p=x_p} \mid A_i)$ are required, where $B_{j=x_j}$ encodes the event where attribute j has the value x_j. Moreover, in order to provide reliable estimates, each possible p-dimensional feature vector $(x_1, \ldots, x_p)$ has to occur in the database sufficiently often.

  By contrast, the estimation of the probabilities under the Naive Bayes Assumption, $P(B_{j=x_j} \mid A_i)$, can be derived from a significantly smaller database since only $p \cdot l$ different "*attribute*$_j$*=x*$_j$"-events $B_{j=x_j}$ are distinguished altogether.

- If the Naive Bayes Assumption applies, then the event A_{NB} will maximize also the posterior probability $P(A_i \mid B_1, \ldots, B_p)$ as defined by the Theorem of Bayes.

- To identify the most probable event, the denominator in the argmax-term, $P(B_1, \ldots, B_p)$, needs not to be estimated since it is constant and cannot influence the ranking among the $\{A_1, \ldots, A_k\}$.

Remarks (continued) :

- Given a set of examples D, then “learning” or “training” a classifier using Naive Bayes means to estimate the prior probabilities (class priors) $P(A_i)$, with $A_i = c(\mathbf{x})$, $(\mathbf{x}, c(\mathbf{x})) \in D$, as well as the probabilities of the observable connections $P(B_{j=x_j} \mid A_i)$, $x_j \in \mathbf{x}$, $j = 1, \ldots, p$, $(\mathbf{x}, c(\mathbf{x})) \in D$, $A_i = c(\mathbf{x})$.

  The obtained probabilities are used in the argmax-term for A_{NB}, which hence encodes the learned hypothesis and functions as a classifier for new feature vectors.

- The hypothesis space H is comprised of all combinations that can be formed from all values that can be chosen for $P(A_i)$ and $P(B_{j=x_j} \mid A_i)$. When building a Naive Bayes classifier, the hypothesis space H is not explored, but the sought hypothesis is directly calculated via a data analysis of D.

- In general the Naive Bayes classifier is not linear, but if the likelihoods, $P($*attribute*$_1$=*x*$_1, \ldots,$ *attribute*$_p$=*x*$_p \mid$ *class*=*c*$_i)$, are from exponential families, the Naive Bayes classifier corresponds to a linear classifier in a particular feature space. [stackexchange]

- The Naive Bayes classifier belongs to the class of *generative models*, which model conditional density functions. By contrast, *discriminative models* attempt to maximize the quality of the output on a training set. [Wikipedia]

Bayes Classification

Naive Bayes (continued)

In addition to the Naive Bayes Assumption, let the following conditions apply:

(c) The set of the k classes is complete: $\sum_{i=1}^{k} P(A_i) = 1,\ \ A_i \in \{c(\mathbf{x}) \mid c(\mathbf{x}) \in D\}$

(d) The A_i are mutually exclusive: $P(A_i, A_\iota) = 0,\ \ 1 \leq i,\ \iota \leq k,\ i \neq \iota$

Bayes Classification

Naive Bayes (continued)

In addition to the Naive Bayes Assumption, let the following conditions apply:

(c) The set of the k classes is complete: $\sum_{i=1}^{k} P(A_i) = 1,\ \ A_i \in \{c(\mathbf{x}) \mid c(\mathbf{x}) \in D\}$

(d) The A_i are mutually exclusive: $P(A_i, A_\iota) = 0,\ \ 1 \leq i,\ \iota \leq k,\ i \neq \iota$

Then holds:

$$P(B_1, \ldots, B_p) \overset{c,d}{=} \sum_{i=1}^{k} P(A_i) \cdot P(B_1, \ldots, B_p \mid A_i) \qquad \text{(theorem of total probability)}$$

$$\overset{NB}{=} \sum_{i=1}^{k} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i) \qquad \text{(Naive Bayes Assumption)}$$

Bayes Classification

Naive Bayes (continued)

In addition to the Naive Bayes Assumption, let the following conditions apply:

(c) The set of the k classes is complete: $\sum_{i=1}^{k} P(A_i) = 1,\ \ A_i \in \{c(\mathbf{x}) \mid c(\mathbf{x}) \in D\}$

(d) The A_i are mutually exclusive: $P(A_i, A_\iota) = 0,\ \ 1 \leq i,\ \iota \leq k,\ \ i \neq \iota$

Then holds:

$$P(B_1, \ldots, B_p) \overset{c,d}{=} \sum_{i=1}^{k} P(A_i) \cdot P(B_1, \ldots, B_p \mid A_i) \quad \text{(theorem of total probability)}$$

$$\overset{NB}{=} \sum_{i=1}^{k} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i) \quad \text{(Naive Bayes Assumption)}$$

With the Theorem of Bayes $(\star)$ it follows for the conditional probabilities:

$$P(A_i \mid B_1, \ldots, B_p) = \frac{P(A_i) \cdot P(B_1, \ldots, B_p \mid A_i)}{P(B_1, \ldots, B_p)} \overset{c,d,NB}{=} \frac{P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i)}{\sum_{i=1}^{k} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i)}$$

Remarks:

- A *ranking* of the $A_1, \ldots, A_k$ can be computed via $\underset{A_i \in \{A_1,\ldots,A_k\}}{\text{argmax}} \; P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i)$.
- If both (c) completeness and (d) mutually exclusiveness of the A_i can be presumed, the total of all posterior probabilities must add up to one: $\sum_{i=1}^{k} P(A_i \mid B_1, \ldots, B_p) = 1$.

  As a consequence, $P(B_1, \ldots, B_p)$ can be estimated and the rank order values for the A_i be "converted" into the respective prior probabilities, $P(A_i \mid B_1, \ldots, B_p)$.

  The normalization is obtained by dividing a rank order value by the rank order values total, $\sum_{i=1}^{k} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i)$.
- The derivation above will in fact yield the true prior probabilities $P(A_i \mid B_1, \ldots, B_p)$, if the Naive Bayes assumption along with the completeness and exclusiveness of the A_i hold.

Bayes Classification

Naive Bayes: Classifier Construction Summary

Let X be a set of feature vectors, C a set of k classes, and $D \subseteq X \times C$ a set of examples. Then the k classes correspond to the events $A_1, \ldots, A_k$, and the p feature values of some $\mathbf{x} \in X$ correspond to the events $B_{1=x_1}, \ldots, B_{p=x_p}$.

Bayes Classification

Naive Bayes: Classifier Construction Summary

Let X be a set of feature vectors, C a set of k classes, and $D \subseteq X \times C$ a set of examples. Then the k classes correspond to the events $A_1, \ldots, A_k$, and the p feature values of some $\mathbf{x} \in X$ correspond to the events $B_{1=x_1}, \ldots, B_{p=x_p}$.

Construction and application of a Naive Bayes classifier:

1. Estimation of the $P(A_i)$, with $A_i = c(\mathbf{x})$, $(\mathbf{x}, c(\mathbf{x})) \in D$.

2. Estimation of the $P(B_{j=x_j} \mid A_i)$, $x_j \in \mathbf{x}$, $j = 1, \ldots, p$, $(\mathbf{x}, c(\mathbf{x})) \in D$, $A_i = c(\mathbf{x})$.

3. Classification of a feature vector $\mathbf{x}$ as A_{NB}, iff

$$A_{NB} = \underset{A_i \in \{A_1, \ldots, A_k\}}{\text{argmax}} \hat{P}(A_i) \cdot \prod_{\substack{x_j \in \mathbf{x} \\ j=1,\ldots,p}} \hat{P}(B_{j=x_j} \mid A_i)$$

4. Given the conditions (c) and (d), computation of the posterior probabilities for A_{NB} as normalization of $\hat{P}(A_{NB}) \cdot \prod_{\substack{x_j \in \mathbf{x} \\ j=1,\ldots,p}} \hat{P}(B_{j=x_j} \mid A_{NB})$.

Remarks:

- There are at most $p \cdot l$ different events $B_{j=x_j}$, if l is an upper bound for the size of the p feature domains.

- The probabilities, denoted as $P(\cdot)$, are unknown and estimated by the relative frequencies, denoted as $\hat{P}(\cdot)$.

- The Naive Bayes approach is adequate for example sets D of medium size up to very large sizes.

- Strictly speaking, the Naive Bayes approach presumes that the feature values in D are "statistically independent given the classes of the target concept". However, experience in the field of text classification shows that convincing classification results are achieved even if the Naive Bayes Assumption does not hold.

- If, in addition to the rank order values, also posterior probabilities shall be computed, both the completeness (c) and the exclusiveness (d) of the target concept classes are required.
  - Requirement (c) is also called *"Closed World Assumption"*.
  - Requirement (d) is also called *"Single Fault Assumption"*.

Bayes Classification

Naive Bayes: Example

	Outlook	Temperature	Humidity	Wind	EnjoySport
1	sunny	hot	high	weak	no
2	sunny	hot	high	strong	no
3	overcast	hot	high	weak	yes
4	rain	mild	high	weak	yes
5	rain	cold	normal	weak	yes
6	rain	cold	normal	strong	no
7	overcast	cold	normal	strong	yes
8	sunny	mild	high	weak	no
9	sunny	cold	normal	weak	yes
10	rain	mild	normal	weak	yes
11	sunny	mild	normal	strong	yes
12	overcast	mild	high	strong	yes
13	overcast	hot	normal	weak	yes
14	rain	mild	high	strong	no

Task: Compute the class $c(\mathbf{x})$ of feature vector $\mathbf{x} = (\textit{sunny}, \textit{cold}, \textit{high}, \textit{strong})$.

Bayes Classification

Naive Bayes: Example (continued)

Let “$B_{j=x_j}$” denotes the event that feature j has value x_j. Then, the feature vector $\mathbf{x} = (\textit{sunny}, \textit{cold}, \textit{high}, \textit{strong})$ gives rise to the following four events:

$B_{1=x_1}$: *Outlook=sunny*

$B_{2=x_2}$: *Temperature=cold*

$B_{3=x_3}$: *Humidity=high*

$B_{4=x_4}$: *Wind=strong*

Bayes Classification

Naive Bayes: Example (continued)

Let "$B_{j=x_j}$" denotes the event that feature j has value x_j. Then, the feature vector $\mathbf{x} = (\textit{sunny}, \textit{cold}, \textit{high}, \textit{strong})$ gives rise to the following four events:

$B_{1=x_1}$: *Outlook=sunny*

$B_{2=x_2}$: *Temperature=cold*

$B_{3=x_3}$: *Humidity=high*

$B_{4=x_4}$: *Wind=strong*

Computation of A_{NB} for $\mathbf{x}$:

$$A_{NB} = \underset{A_i \in \{\textit{ yes, no }\}}{\text{argmax}} \quad \hat{P}(A_i) \cdot \prod_{\substack{x_j \in \mathbf{x} \\ j=1,\ldots,4}} \hat{P}(B_{j=x_j} \mid A_i)$$

$$= \underset{A_i \in \{\textit{ yes, no }\}}{\text{argmax}} \quad \hat{P}(A_i) \cdot \hat{P}(\textit{Outlook=sunny} \mid A_i) \cdot \hat{P}(\textit{Temperature=cold} \mid A_i) \cdot \hat{P}(\textit{Humidity=high} \mid A_i) \cdot \hat{P}(\textit{Wind=strong} \mid A_i)$$

Bayes Classification

Naive Bayes: Example (continued)

To classify $\mathbf{x}$ altogether $2+4\cdot 2$ probabilities are estimated from the data:

- ❑ $\hat{P}(\textit{EnjoySport=yes}) = \frac{9}{14} = 0.64$
- ❑ $\hat{P}(\textit{EnjoySport=no}) = \frac{5}{14} = 0.36$
- ❑ $\hat{P}(\textit{Wind=strong} \mid \textit{EnjoySport=yes}) = \frac{3}{9} = 0.33$
- ❑ $\ldots$

Bayes Classification

Naive Bayes: Example (continued)

To classify $\mathbf{x}$ altogether $2 + 4 \cdot 2$ probabilities are estimated from the data:

- $\hat{P}(\textit{EnjoySport=yes}) = \frac{9}{14} = 0.64$
- $\hat{P}(\textit{EnjoySport=no}) = \frac{5}{14} = 0.36$
- $\hat{P}(\textit{Wind=strong} \mid \textit{EnjoySport=yes}) = \frac{3}{9} = 0.33$
- ...

➜ Ranking:

1. $\hat{P}(\textit{EnjoySport=no}) \cdot \prod_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=no}) = 0.0206$
2. $\hat{P}(\textit{EnjoySport=yes}) \cdot \prod_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=yes}) = 0.0053$

Bayes Classification

Naive Bayes: Example (continued)

To classify $\mathbf{x}$ altogether $2 + 4 \cdot 2$ probabilities are estimated from the data:

- $\hat{P}(\textit{EnjoySport=yes}) = \frac{9}{14} = 0.64$
- $\hat{P}(\textit{EnjoySport=no}) = \frac{5}{14} = 0.36$
- $\hat{P}(\textit{Wind=strong} \mid \textit{EnjoySport=yes}) = \frac{3}{9} = 0.33$
- $\ldots$

➜ Ranking:

1. $\hat{P}(\textit{EnjoySport=no}) \cdot \prod_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=no}) = 0.0206$
2. $\hat{P}(\textit{EnjoySport=yes}) \cdot \prod_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=yes}) = 0.0053$

➜ Probabilities: (subject to conditions (c) and (d))

1. $\hat{P}(\textit{EnjoySport=no} \mid \mathbf{x}) = \frac{0.0206}{0.0053+0.0206} \approx 80\%$
2. $\hat{P}(\textit{EnjoySport=yes} \mid \mathbf{x}) = \frac{0.0053}{0.0053+0.0206} \approx 20\%$