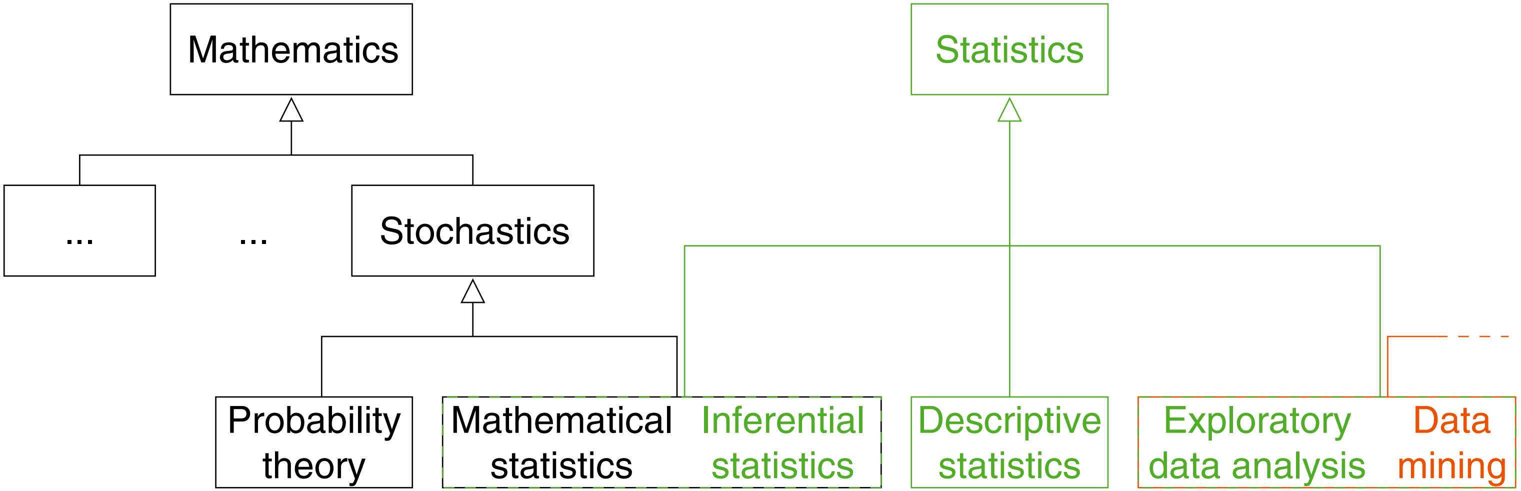
Chapter ML:IV

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 equivalent different systems.
- 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. As a consequence, a random experiment may turn to a deterministic process if 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.

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.	$\emptyset$		impossible event
2.	Ω		certain event
3.	$\overline{A} := \Omega \setminus A$		complementary event (opposite event) of A
4.	$\|A\| = 1$		elementary event
5.	$A \subseteq B$	$\Leftrightarrow$	A is a sub-event of B or "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 (compatible otherwise)

Probability Basics

Classical Concept Formation

Empirical law of large numbers:

For particular events 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:

For particular events 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, 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, one has tried to develop a frequentist probability concept that is based on the (fictitious) limit of the relative frequencies [von Mises, 1951]. The attempt failed since such a limit formation is only within mathematical settings possible (infinitesimal calculus), where accurate repetitions unto infinity can be made.

Probability Basics

Axiomatic Concept Formation

The principle steps of axiomatic concept formation:

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

2. Specify the basic, required properties of this function in the form of axioms.

Probability Basics

Axiomatic Concept Formation

The principle steps of axiomatic concept formation:

1. Postulate a function that assigns a probability to each element of the event space.
2. Specify the basic, 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. Then a function $P : \mathcal{P}(\Omega) \rightarrow \mathbf{R}$ that 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 (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 the distribution of the probabilities P.
- Generally, a function that is equipped with 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 $A, B \in \mathcal{P}(\Omega)$ 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 event equals the sum of the probabilities of the sub-events into which the event has been partitioned.

- Considered as a function in parameter A and 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: $P_B(\overline{A}) = 1 - P_B(A)$

- The following inequality must usually be assumed: $P(A \mid \overline{B}) \neq 1 - P(A \mid B)$.

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 $2 \leq l \leq k$.

Chapter ML:IV (continued)

IV. Statistical Learning

- ❑ Probability Basics
- ❑ Bayes Classification
- ❑ Maximum a-Posteriori Hypotheses

Bayes Classification

Theorem 12 (Bayes)

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:

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

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

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

Bayes Classification

Theorem 12 (Bayes)

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:

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

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

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

Proof

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 the total probability for $P(B)$ in the denominator will yield the claim of the theorem.

Bayes Classification

Combined Conditions

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

Applied to a classification problem:

- A corresponds to an event of kind “class=c”, and the B_j, $j = 1, \ldots, p$, correspond to p events of kind “attribute=value”.
- observable connection (standard situation) : $B_1, \ldots, B_p \mid A$
- reversed connection (diagnosis situation) : $A \mid B_1, \ldots, B_p$

Bayes Classification

Combined Conditions

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

Applied to a classification problem:

- ❑ A corresponds to an event of kind “class=c”, and the B_j, $j = 1, \ldots, p$, correspond to p events of kind “attribute=value”.
- ❑ observable connection (standard situation) : $B_1, \ldots, B_p \mid A$
- ❑ reversed connection (diagnosis situation) : $A \mid B_1, \ldots, B_p$

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

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

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 (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 $B_1, \ldots, B_p$ occur, and one is interested in the likelihood that A is given or has been occurred.
  - Based on the prior probabilities of the classes (aka class priors), $P($class=c$)$, and the probabilities of the observable connections, $P($attribute=value $|$ class=c$)$, the conditional class probabilities in an analysis situation, $P($class=c $|$ attribute=value$)$, can be computed with the theorem of Bayes.

- The class-conditional event “attribute=value | class=c” does not necessarily model a cause-effect relation: the event “class=c” *may* cause—but does not need to cause—the event “attribute=value”.

Remarks (continued) :

- $P(A \mid B_1, \ldots, B_p)$ is called conditional probability of A given the conditions $B_1, \ldots, B_p$.
- Alternative and semantically equivalent notations of $P(A \mid B_1, \ldots, B_p)$ are:
  1. $P(A \mid B_1, \ldots, B_p)$
  2. $P(A \mid B_1 \land \ldots \land B_p)$
  3. $P(A \mid B_1 \cap \ldots \cap B_p)$

Bayes Classification

Naive Bayes

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

(a) Naive Bayes Assumption: “Given condition A, the $B_1, \ldots, B_p$ are statistically independent” (aka: the B_i are conditionally independent). Formally:

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

Bayes Classification

Naive Bayes

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

(a) Naive Bayes Assumption: “Given condition A, the $B_1, \ldots, B_p$ are statistically independent” (aka: the B_i are conditionally independent). Formally:

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

(b) $P(B_1, \ldots, B_p)$ is constant and hence needs not to be estimated if one is interested only in the most likely event under the Naive Bayes Assumption, $A_{NB} \in \{A_1, \ldots, A_k\}$. A_{NB} can be computed with the theorem of Bayes $(\star)$:

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

Remarks:

- Why the probabilities $P(B_1, \ldots, B_p \mid A)$ usually cannot be estimated in the wild: Suppose that we are given k classes, and that the domains of the p attributes of a feature vector contain minimum l values each, then for as many as $k \cdot p^l$ different feature vectors (= class-features-values combinations) the probability values are required. In order to provide reliable estimates, each class-features-values combination must occur in the database sufficiently frequently. By contrast, the estimation of the probabilities $P(B \mid A)$ can be derived from a significantly smaller database since only $p \cdot l \cdot k$ combined events are distinguished altogether.

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

- Given a set of examples D, then “learning” or “training” a classifier using Naive Bayes means to estimate the prior probabilities (class priors) $P(A)$, where $A \in \{c(\mathbf{x}) \mid (\mathbf{x}, c(\mathbf{x})) \in D\}$, as well as the probabilities of the observable connections $P(B \mid A)$, where $B \in \{B_{j=x_j} \mid x_j \in \mathbf{x}, (\mathbf{x}, c(\mathbf{x})) \in D\}$ and $A = c(\mathbf{x})$. 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 $P(A)$ and $P(B \mid A)$. When constructing a Naive Bayes classifier, the hypothesis space H is not explored, but the sought hypothesis is directly computed from a data analysis of D.

  Keyword: *discriminative* classifier versus *generative* classifier

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:

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

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

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 $\underset{A \in \{A_1,\ldots,A_k\}}{\text{argmax}} \; P(A) \cdot \prod_{j=1}^{p} P(B_j \mid A)$.
- 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, the rank order values of the A_i can be “converted into the 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, i.e., $\sum_{i=1}^{k} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i)$.
- 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 p-dimensional feature space, let C be the set of k classes of a target concept, and let D be a set of examples of the form $(\mathbf{x}, c(\mathbf{x}))$ over $X \times C$. 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 p-dimensional feature space, let C be the set of k classes of a target concept, and let D be a set of examples of the form $(\mathbf{x}, c(\mathbf{x}))$ over $X \times C$. 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 $P(A)$, where $A = c(\mathbf{x})$, $(\mathbf{x}, c(\mathbf{x})) \in D$.

2. Estimation of the $P(B_{j=x_j} \mid A)$, where $x_j \in \mathbf{x}$, $(\mathbf{x}, c(\mathbf{x})) \in D$, $c(\mathbf{x}) = A$.

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

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

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 a very large size.
- 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. The first requirement is also called *"Closed World Assumption"*, the second requirement 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

Let the target concept $c(\mathbf{x})$ of feature vector $\mathbf{x} = (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})$ be unknown.

Bayes Classification

Naive Bayes: Example (continued)

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

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

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

Bayes Classification

Naive Bayes: Example (continued)

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

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

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

"$B_{j=x_j}$" denotes the event for a particular attribute-value-combination in $\mathbf{x}$, namely, that event where attribute (dimension) j has value x_j.

The feature vector $\mathbf{x} = (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})$ with the unknown target concept gives rise to the following four events:

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

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

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

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

Bayes Classification

Naive Bayes: Example (continued)

For the classification of $\mathbf{x}$ altogether $2 + 4 \cdot 2$ probabilities have to be estimated:

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

Bayes Classification

Naive Bayes: Example (continued)

For the classification of $\mathbf{x}$ altogether $2 + 4 \cdot 2$ probabilities have to be estimated:

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

➜ Ranking:

1. $\hat{P}(\textit{EnjoySport=no}) \cdot \prod\limits_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=no}) = 0.0206$

2. $\hat{P}(\textit{EnjoySport=yes}) \cdot \prod\limits_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=yes}) = 0.0053$

Bayes Classification

Naive Bayes: Example (continued)

For the classification of $\mathbf{x}$ altogether $2 + 4 \cdot 2$ probabilities have to be estimated:

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

➜ 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. $\hat{P}(\textit{EnjoySport=yes}) \cdot \prod_{x_j \in \mathbf{x}} \hat{P}(B_{j=x_j} \mid \textit{EnjoySport=yes}) = 0.0053$

➜ Normalization: (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. $\hat{P}(\textit{EnjoySport=yes} \mid \mathbf{x}) = \frac{0.0053}{0.0053+0.0206} \approx 20\%$