# Chapter ML:VII (continued)

## VII. Bayesian Learning

- Approaches to Probability
- Conditional Probability
- Bayes Classifier
- Exploitation of Data
- Frequentist versus Subjectivist

# Bayes Classifier

## Generative Approach to Classification Problems

Setting:

- $X$ is a multiset of feature vectors.
- $C$ is a set of classes.
- $D = \{(\mathbf{x}_1, c_1), \ldots, (\mathbf{x}_n, c_n)\} \subseteq X \times C$ is a multiset of examples.

Learning task:

- Fit $D$ using joint probabilities $p()$ between features and classes.

# Bayes Classifier

## 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:

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

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

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

# Bayes Classifier

## 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 Classifier

## 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 Classifier

## Example: Reasoning About a Disease [Kirchgessner 2009]

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

   $B$ : *test_pos*

# Bayes Classifier

## 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 Classifier

## 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*

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 Classifier

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

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

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

$B$ : *test_pos*

2. $B \mid A_1$ : *test_pos* | *Aids* $\quad P(B \mid A_1) = 0.98$ (result from clinical trials)
3. $B \mid A_2$ : *test_pos* | *no_Aids* $\quad P(B \mid A_2) = 0.03$ (result from clinical trials)

# Bayes Classifier

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

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

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

$B$ : *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* $\quad P(B \mid A_1) = 0.98$ (result from clinical trials)
3. $B \mid A_2$ : *test_pos* | *no_Aids* $\quad P(B \mid A_2) = 0.03$ (result from clinical trials)

# Bayes Classifier

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

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

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

$B$ : *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* $\quad P(B \mid A_1) = 0.98$ (result from clinical trials)
3. $B \mid A_2$ : *test_pos* | *no_Aids* $\quad P(B \mid A_2) = 0.03$ (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 Classifier

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

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

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

$B$ : *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* $\quad P(B \mid A_1) = 0.98$ (result from clinical trials)
3. $B \mid A_2$ : *test_pos* | *no_Aids* $\quad P(B \mid A_2) = 0.03$ (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 Classifier

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

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

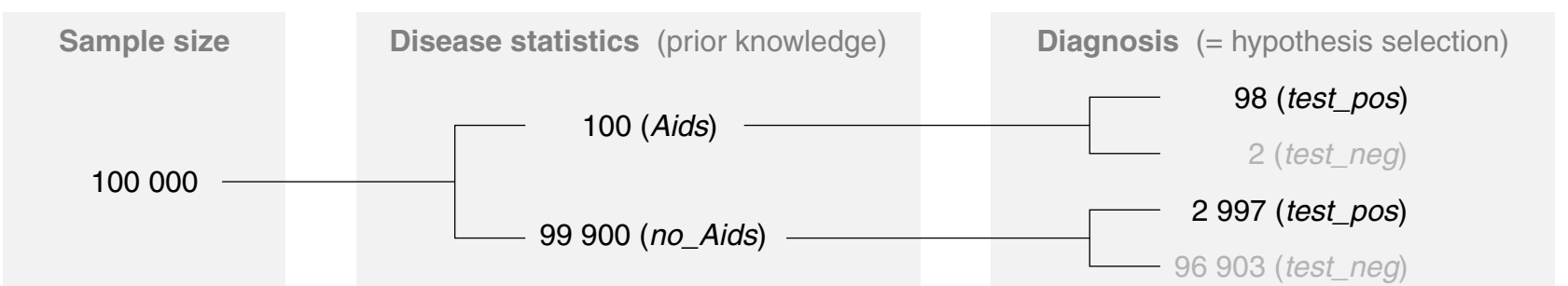
$\Rightarrow$ $A_2$ : *no_Aids* $\quad P(A_2) = P(\overline{A_1}) = 1 - P(A_1) = 0.999$

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

2. $B \mid A_1$ : *test_pos* | *Aids* $\quad P(B \mid A_1) = 0.98$ (result from clinical trials)
3. $B \mid A_2$ : *test_pos* | *no_Aids* $\quad P(B \mid A_2) = 0.03$ (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 Classifier

## 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 Classifier

## 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 "$C{=}c_i$",
  the $B_j$, $j = 1, \ldots, p$, correspond to $p$ events of the kind "$X_j{=}x_j$".
- observable relation (in the prevalent setting) : $B_1, \ldots, B_p \mid A_i$
- reversed relation (in a diagnosis setting) : $A_i \mid B_1, \ldots, B_p$

# Bayes Classifier

## 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 "$C{=}c_i$",
  the $B_j$, $j = 1, \ldots, p$, correspond to $p$ events of the kind "$X_j{=}x_j$".
- observable relation (in the prevalent setting) : $B_1, \ldots, B_p \mid A_i$
- reversed relation (in a diagnosis setting) : $A_i \mid B_1, \ldots, B_p$

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 feature-value pairs are interpreted as events. The relation to an underlying sample space $\Omega$, $\Omega = \{\omega_1, \ldots, \omega_n\}$, from which the events are subsets, is not considered.
  - Observable or measurable and possibly causal relation: 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 reasoning".
  - "Backward reasoning", 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), $P(C{=}c_i)$, and the class-conditional probabilities of the observable relations (aka likelihoods), $P(X_1{=}x_1, \ldots, X_p{=}x_p \mid C{=}c_i)$, the conditional class probabilities in an analysis situation, $P(C{=}c_i \mid X_1{=}x_1, \ldots, X_p{=}x_p)$, can be computed with the Theorem of Bayes.
- The $X_j$ and $C$ denote random variables with ranges of the respective feature domains and $C$ respectively.
- A conditional event "$X_j{=}x_j \mid C{=}c_i$" does not necessarily model a cause-effect relation: the event "$C{=}c_i$" may cause—but does not need to cause—the event "$X_j{=}x_j$".

Remarks: (continued)

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

## 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:

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

# Bayes Classifier

## 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:

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

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

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

# Bayes Classifier

## 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:

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

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

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

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 $p$ features and that the domains of the features contain minimum $m$ values each.

  Then, for as many as $m^p$ 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 feature $j$ has 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 $p \cdot m$ different "$X_j{=}x_j$"-events $B_{j=x_j}$ are distinguished altogether.

- If the Naive Bayes Assumption applies, then the event $A_{\mathsf{NB}}$ 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: it is constant and has no influence on the ranking among the $\{A_1, \ldots, A_k\}$. See the following identities:

$$\begin{aligned} \operatorname{argmax}_{A_i} P(A_i \mid B) &= \operatorname{argmax}_{A_i} \frac{P(B|A_i) \cdot P(A_i)}{P(B)} && \text{Bayes rule.} \\ &= \operatorname{argmax}_{A_i} P(B \mid A_i) \cdot P(A_i) && P(B) \text{ does not depend on } A_i. \end{aligned}$$

Remarks: (continued)

- Given a multiset of examples $D$, then "learning" or "training" a classifier via Naive Bayes means to estimate the prior probabilities (class priors) $P(A_i)$, with $A_i := C{=}c_i$, $i = 1, \ldots, k$, as well as the probabilities of the observable relations $P(B_{j=x_j} \mid A_i)$, with $B_{j=x_j} := X_j{=}x_j$, $j = 1, \ldots, p$.

  These probabilities are used in the argmax-term for $A_{\text{NB}}$, which encodes the "learned" hypothesis and functions as a classifier for new feature vectors.

- The hypothesis space $H$ is the space of candidate target functions that exploit the probabilities $P(A_i)$ and $P(B_{j=x_j} \mid A_i)$ to map from a set of "feature events", each denoted as $B_{j=x_j}$, onto a "class event", denoted as $A_i$. Under Naive Bayes the hypothesis space $H$ is not explored, but the sought hypothesis is directly constructed as $A_{\text{NB}}$.

- In general, the Naive Bayes classifier is not linear in the sense that the generated decision boundary in the input space is non-linear (= is not a hyperplane). If the likelihoods, $P(X_1{=}x_1, \ldots, X_p{=}x_p \mid C{=}c_i)$, 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 probability mass (or density) functions. By contrast, *discriminative models* minimize the loss or misclassification error. [Wikipedia]

# Bayes Classifier

## Naive Bayes (continued)

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

(c) The set of the $k$ classes is complete: $\sum_{i=1}^{k} P(A_i) = 1$

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

# Bayes Classifier

## Naive Bayes (continued)

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

(c) The set of the $k$ classes is complete: $\sum_{i=1}^{k} P(A_i) = 1$

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

Then holds:

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

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

# Bayes Classifier

## Naive Bayes (continued)

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

(c) The set of the $k$ classes is complete: $\sum_{i=1}^{k} P(A_i) = 1$

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

Then holds:

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

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

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

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

Remarks:

- A *ranking* of the $A_1, \ldots, A_k$ can be computed via $\underset{A_i \in \{A_1,\ldots,A_k\}}{\operatorname{argmax}} P(A_i) \cdot \prod_{j=1}^{p} P(B_j \mid A_i)$.

- 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 Classifier

## Naive Bayes: Classifier Construction Summary

Let $X$ be a multiset of feature vectors, $C$ a set of $k$ classes, and $D \subseteq X \times C$ a multiset 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 Classifier

## Naive Bayes: Classifier Construction Summary

Let $X$ be a multiset of feature vectors, $C$ a set of $k$ classes, and $D \subseteq X \times C$ a multiset 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. Using $D$, estimate the $P(A_i)$, $A_i := C{=}c_i$, $i = 1, \ldots, k$.
2. Using $D$, estimate the $P(B_{j=x_j} \mid A_i)$, $B_{j=x_j} := X_j{=}x_j$, $j = 1, \ldots, p$.
3. Classify feature vector $\mathbf{x}$ as $A_{\text{NB}}$, iff

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

4. Given (c) and (d), estimate the posterior probabilities $P(A_i \mid B_1, \ldots, B_p)$ by normalizing $\hat{P}(A_i) \cdot \prod_{\substack{x_j \in \mathbf{x} \\ j=1,\ldots,p}} \hat{P}(B_{j=x_j} \mid A_i)$ with $\sum_{i=1}^{k} \hat{P}(A_i) \cdot \prod_{j=1}^{p} \hat{P}(B_j \mid A_i)$

Remarks:

- There are at most $p \cdot m$ different events $B_{j=x_j}$, if $m$ is an upper bound for the size of the $p$ feature domains.
- Recap. The probabilities, denoted as $P()$, are unknown and are estimated by the relative frequencies, denoted as $\hat{P}()$.
- 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 Classifier

## Naive Bayes: Example

A multiset of examples $D$:

| | Outlook | Temperature | Humidity | Wind | EnjoySurfing |
|---|---|---|---|---|---|
| 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 |

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

# Bayes Classifier

## Naive Bayes: Example (continued)

Let "$B_{j=x_j}$" denotes the event that feature $j$ has value $x_j$. Then, the feature vector $\mathbf{x} = (\text{sunny}, \text{cold}, \text{high}, \text{strong})$ 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 Classifier

## Naive Bayes: Example (continued)

Let "$B_{j=x_j}$" denotes the event that feature $j$ has value $x_j$. Then, the feature vector $\mathbf{x} = (\text{sunny}, \text{cold}, \text{high}, \text{strong})$ 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 $A_{\text{NB}}$ for $\mathbf{x}$ :

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

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

# Bayes Classifier

## Naive Bayes: Example (continued)

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

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

# Bayes Classifier

## Naive Bayes: Example (continued)

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

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

➜ Ranking:

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

# Bayes Classifier

## Naive Bayes: Example (continued)

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

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

➜ Ranking:

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

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

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

Remarks:

- **X** denotes a four-dimensional random variable (a random vector) with possible realizations as defined in the data $D$.