Chapter ML:IV (continued)

IV. Statistical Learning

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

Bayes Classification

Single Conditional Event

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

Single Conditional Event

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 $P(B)$ will yield the claim.

Bayes Classification

Combined Conditional Events

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

Applied to a classification problem:

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

Bayes Classification

Combined Conditional Events

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

Applied to a classification problem:

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

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

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

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 $B_1, \ldots, B_m$ occur. One may denote this as forward connection.
  - Reversed connection, typically an analysis or diagnosis situation: the symptoms $B_1, \ldots, B_m$ occur, and one is interested in the likelihood that A is given or has occurred.
  - Based on the prior probabilities of the classes (aka class priors), $P(\text{class=c})$, and the probabilities of the observable connections (aka likelihoods), $P(\text{attribute=value} \mid \text{class=c})$, the conditional class probabilities in an analysis situation, $P(\text{class=c} \mid \text{attribute=value})$, 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):

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

Bayes Classification

Naive Bayes

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

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

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

Bayes Classification

Naive Bayes

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

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

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

(b) $P(B_1, \ldots, B_m)$ 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 $(\star)$:

$$\operatorname*{argmax}_{A \in \{A_1,\ldots,A_k\}} \frac{P(A) \cdot P(B_1, \ldots, B_m \mid A)}{P(B_1, \ldots, B_m)} \overset{NB}{=} \operatorname*{argmax}_{A \in \{A_1,\ldots,A_k\}} P(A) \cdot \prod_{j=1}^{m} P(B_j \mid A) \quad = \quad A_{NB}$$

Remarks:

- Usually the probability $P(B_1, \ldots, B_m \mid A)$ 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 $P(B_{1_l}, \ldots, B_{p_l} \mid A)$ are required. In order to provide reliable estimates, each possible p-dimensional feature vector $(x_1, \ldots, x_p)$ must occur in the database sufficiently often. By contrast, the estimation of the probabilities $P(B \mid A)$ can be derived from a significantly smaller database since only $p \cdot l$ "attribute=value events" B are distinguished altogether.
- If the Naive Bayes Assumption applies, then the event A_{NB} will maximize also the posterior probability $P(A \mid B_1, \ldots, B_m)$ 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 $B \in \{B_{j=x_j} \mid j = 1, \ldots, p,\ x_j \in \mathbf{x},\ (\mathbf{x}, c(\mathbf{x})) \in D\}$ 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:

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

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

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:

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

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

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

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

Remarks:

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

- 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: $\sum_{i=1}^{k} P(A_i \mid B_1, \ldots, B_m) = 1$.

  As a consequence, the rank order values of the A_i can be "converted into the prior probabilities" $P(A_i \mid B_1, \ldots, B_m)$. The normalization is obtained by dividing a rank order value by the rank order values total, $\sum_{i=1}^{k} P(A_i) \cdot \prod_{j=1}^{m} P(B_j \mid A_i)$.

- The derivation above will in fact yield the true prior probabilities $P(A_i \mid B_1, \ldots, B_m)$, 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 $j = 1, \ldots, p$, $x_j \in \mathbf{x}$, $(\mathbf{x}, c(\mathbf{x})) \in D$, $A = c(\mathbf{x})$.
3. Classification of a feature vector $\mathbf{x}$ as A_{NB}, iff

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

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 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. 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

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)

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

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

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

Bayes Classification

Naive Bayes: Example (continued)

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

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

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

"$B_{j=x_j}$" denotes the event that feature (dimension) j has value x_j.

The feature vector $\mathbf{x} = (\textit{sunny}, \textit{cold}, \textit{high}, \textit{strong})$ with the unknown class 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)

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

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

Bayes Classification

Naive Bayes: Example (continued)

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

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

➜ 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)

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

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

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