Bayesian Learning: Introduction

Decision Making

Bayes Theorem allows us to update our beliefs / hypothesis H based on new evidence / data D. It is a way of thinking with uncertainty.

$$P[H \mid D] = \frac{P[H] \cdot P[D \mid H]}{P[D]}$$

Bayesian Learning: Introduction

Decision Making

Bayes Theorem allows us to update our beliefs / hypothesis H based on new evidence / data D. It is a way of thinking with uncertainty.

$$P[H \mid D] = \frac{P[H] \cdot P[D \mid H]}{P[D]}$$

Example: We get headache and coughing as symptoms. Using Google we know that 90% of the people having the flue, are showing symptoms of headache and coughing. We also know that 5% of the humans are getting the flue and that headache and coughing occurs in 20% of the humans. How likely is it that we have the flue?

Bayesian Learning: Introduction

Decision Making

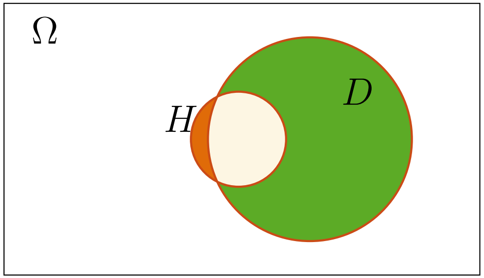
Bayes Theorem allows us to update our beliefs / hypothesis H based on new evidence / data D. It is a way of thinking with uncertainty.

$$P[H \mid D] = \frac{P[H] \cdot P[D \mid H]}{P[D]}$$

Example: We get headache and coughing as symptoms. Using Google we know that 90% of the people having the flue, are showing symptoms of headache and coughing. We also know that 5% of the humans are getting the flue and that headache and coughing occurs in 20% of the humans. How likely is it that we have the flue?

- H = We have the flue
- D = Headache and Coughing

$$\begin{aligned} P[H \mid D] &= \frac{P[H] \cdot P[D \mid H]}{P[D]} \\ &= \frac{0.05 \cdot 0.9}{0.2} = 0.225 = 22.5\% \end{aligned}$$



Bayesian Learning: Introduction

Decision Making

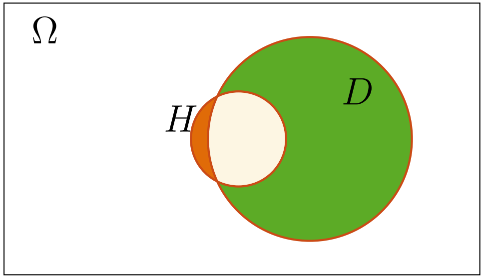
Bayes Theorem allows us to update our beliefs / hypothesis H based on new evidence / data D. It is a way of thinking with uncertainty.

$$P[H \mid D] = \frac{P[H] \cdot P[D \mid H]}{P[D]}$$

Example: We get headache and coughing as symptoms. Using Google we know that 90% of the people having the flue, are showing symptoms of headache and coughing. We also know that 5% of the humans are getting the flue and that headache and coughing occurs in 20% of the humans. How likely is it that we have the flue?

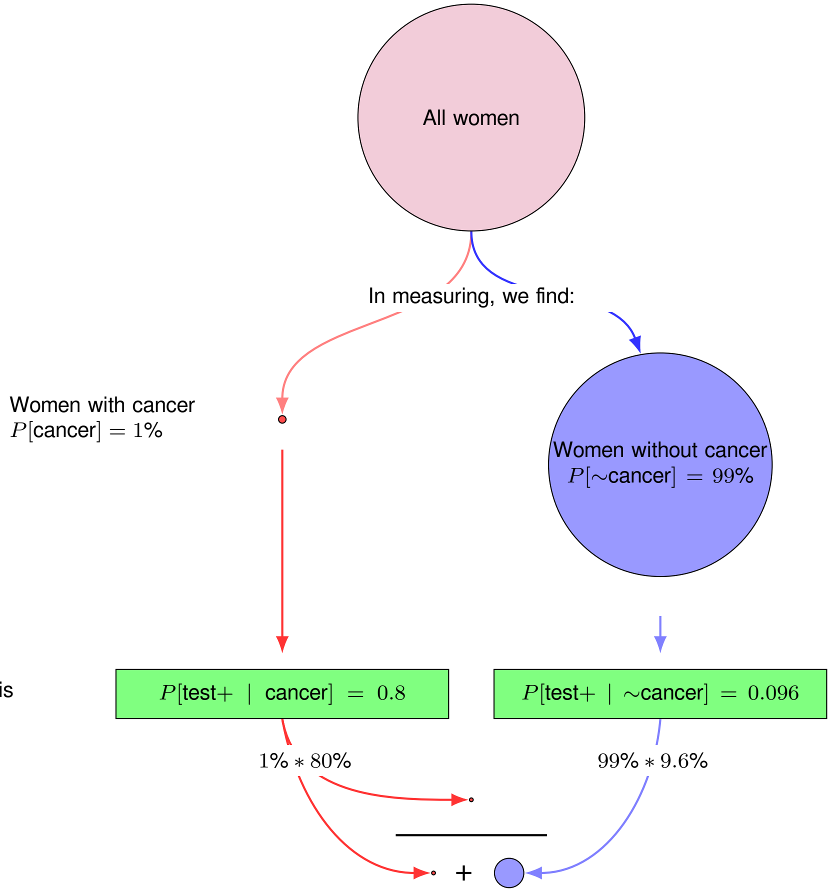
- H = We have the flue
- D = Headache and Coughing

$$\begin{aligned} P[H \mid D] &= \frac{P[H] \cdot P[D \mid H]}{P[D]} \\ &= \frac{0.05 \cdot 0.9}{0.2} = 0.225 = 22.5\% \end{aligned}$$



Two Reasons: Only a few people have the flue and the symptoms are occuring more frequently!

Bayesian Learning: Introduction
Graphical Illustration



Now we pass both groups through the sieve; note that both sieves are *the same*; they just behave differently depending on which group is passing through.
Let test+ = a positve mammography.

Finally, to find the probability that a positive test *actually means cancer*, we look at those who passed through the sieve *with cancer*, and divide by all who received a positive test, cancer or not.

$$\frac{P[(]\text{test+} \mid \text{cancer})}{P[(]\text{test+} \mid \text{cancer}) + P[(]\text{test+} \mid \sim\text{cancer})} =$$

$$\frac{1\% * 80\%}{(1\% * 80\%) + (99\% * 9.6\%)} = 7.8\% = P[(]\text{cancer} \mid \text{test+})$$ [1]

[1] Source: http://www.texample.net/tikz/examples/bayes/

Bayesian Learning: Principle

Applying Bayes Theorem to Classification

Let $P[B \mid A_1, \ldots, A_p]$ denote the probability of the occurrence of event B given that the events $A_1, \ldots, A_p$ are known to have occurred.

Applied to a classification problem, i.e. whether certain data provides evidence that the data falls in a particular class:

- the A_j, $j = 1, \ldots, p$, correspond to p events of type “attribute=value”, B corresponds to an event of type “class=y”.
- observed connection (standard situation) : $A_1, \ldots, A_p \mid B$
- reversed connection (diagnosis situation) : $B \mid A_1, \ldots, A_p$

Bayesian Learning: Principle

Applying Bayes Theorem to Classification

Let $P[B \mid A_1, \ldots, A_p]$ denote the probability of the occurrence of event B given that the events $A_1, \ldots, A_p$ are known to have occurred.

Applied to a classification problem, i.e. whether certain data provides evidence that the data falls in a particular class:

- the A_j, $j = 1, \ldots, p$, correspond to p events of type "attribute=value", B corresponds to an event of type "class=y".
- observed connection (standard situation) : $A_1, \ldots, A_p \mid B$
- reversed connection (diagnosis situation) : $B \mid A_1, \ldots, A_p$

If sufficient data for estimating $P[B]$ and $P[A_1, \ldots, A_p \mid B]$ is provided, then $P[B \mid A_1, \ldots, A_p]$ can be computed with the theorem of Bayes:

$$P[B \mid A_1, \ldots, A_p] = \frac{P[B] \cdot P[A_1, \ldots, A_p \mid B]}{P[A_1, \ldots, A_p]} \qquad (\star)$$

Remarks:

- $P[B \mid A_1, \ldots, A_p]$ is called conditional probability of B given the conditions $A_1, \ldots, A_p$. Alternative, semantically equivalent notations are:

  1. $P[B \mid A_1, \ldots, A_p]$
  2. $P[B \mid A_1 \wedge \ldots \wedge A_p]$
  3. $P[B \mid A_1 \cap \ldots \cap A_p]$

- 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 $\Omega = \{\omega_1, \ldots, \omega_n\}$, from which the events are subsets, is not considered.

  – Observable or measurable, possibly causal connection: it is (or was in the past) regularly observed that in situation B the symptoms $A_1, \ldots, A_p$ occur. One may denote this as forward connection.

  – Reversed connection, typically an analysis or diagnosis situation: the symptoms $A_1, \ldots, A_p$ occur, and one is interested in the likelihood that B is given or has been occurred.

  – Based on the prior probabilities of the classes (aka class priors), $P[$class=y$]$, and the probabilities of the observable connections, $P[$attribute=value $\mid$ class=y$]$, the conditional class probabilities in an analysis situation, $P[$class=c $\mid$ attribute=value$]$, can be computed with the theorem of Bayes.

- The class-conditional event “attribute=value $\mid$ 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”.

Bayesian Learning: Naive Bayes

Naive Bayes

The compilation of a database from which reliable values for the $P[A_1, \ldots, A_p \mid B]$ can be obtained is often infeasible. The way out:

(a) Naive Bayes Assumption: "Given B, the $A_1, \ldots, A_p$ are statistically independent" (aka conditional independence). Formally:

$$P[A_1, \ldots, A_p \mid B] \stackrel{NB}{=} \prod_{j=1}^{p} P[A_j \mid B]$$

Bayesian Learning: Naive Bayes

Naive Bayes

The compilation of a database from which reliable values for the $P[A_1, \ldots, A_p \mid B]$ can be obtained is often infeasible. The way out:

(a) Naive Bayes Assumption: "Given B, the $A_1, \ldots, A_p$ are statistically independent" (aka conditional independence). Formally:

$$P[A_1, \ldots, A_p \mid B] \stackrel{NB}{=} \prod_{j=1}^{p} P[A_j \mid B]$$

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

$$\underset{B \in \{B_1, \ldots, B_k\}}{\text{argmax}} \frac{P[B] \cdot P[A_1, \ldots, A_p \mid B]}{P[A_1, \ldots, A_p]} \stackrel{NB}{=} \underset{B \in \{B_1, \ldots, B_k\}}{\text{argmax}} P[B] \cdot \prod_{j=1}^{p} P[A_j \mid B] = B_{NB}$$

Remarks:

- Why the probabilities $P[A_1, \ldots, A_p \mid B]$ can usually not 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 $P[A \mid B]$ can be derived from a significant smaller database since only $k \cdot p \cdot l$ combined events are distinguished altogether.

- If the Naive Bayes Assumption applies, then the event B_{NB} will maximize also the posterior probability $P[B \mid A_1, \ldots, A_p]$ 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) $P[B]$, where $B \in \{y(\mathbf{x}) \mid (\mathbf{x}, y(\mathbf{x})) \in D\}$, as well as the probabilities of the observable connections $P[A \mid B]$, where $A \in \{A_{j=x_j} \mid x_j \in \mathbf{x}, (\mathbf{x}, y(\mathbf{x})) \in D\}$ and $y(\mathbf{x}) = B$. The obtained probabilities are used in the optimization term for B_{NB}, which hence encodes the learned hypothesis and functions as a classifier for new feature vectors.

- The hypothesis space H comprises all values that can be chosen for $P[B]$ and $P[A \mid B]$. When constructing a Naive Bayes classifier, the hypothesis space H is not explored, but the sought hypothesis is directly computed from an descriptive data analysis of D.

  Keyword: *discriminative* classifier versus *generative* classifier

Bayesian Learning: Naive Bayes

Naive Bayes (continued)

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

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

(d) the B_i are mutually exclusive: $P[B_i, B_\iota] = 0$, $1 \leq i,\ \iota \leq k,\ i \neq \iota$

Bayesian Learning: Naive Bayes

Naive Bayes (continued)

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

(c) the set of the k classes is complete: $\sum_{i=1}^{k} P[B_i] = 1,\; B_i \in \{y(\mathbf{x}) \mid y(\mathbf{x}) \in D\}$

(d) the B_i are mutually exclusive: $P[B_i, B_\iota] = 0,\; 1 \leq i,\; \iota \leq k,\; i \neq \iota$

Then holds:

$$P[A_1, \ldots, A_p] = \sum_{i=1}^{k} P[B_i] \cdot P[A_1, \ldots, A_p \mid B_i] \quad \text{(theorem of total probability)}$$

$$\stackrel{NB}{=} \sum_{i=1}^{k} P[B_i] \cdot \prod_{j=1}^{p} P[A_j \mid B_i] \quad \text{(Naive Bayes Assumption)}$$

Bayesian Learning: Naive Bayes

Naive Bayes (continued)

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

(c) the set of the k classes is complete: $\sum_{i=1}^{k} P[B_i] = 1, \;\; B_i \in \{y(\mathbf{x}) \mid y(\mathbf{x}) \in D\}$

(d) the B_i are mutually exclusive: $P[B_i, B_\iota] = 0, \;\; 1 \leq i, \; \iota \leq k, \; i \neq \iota$

Then holds:

$$P[A_1, \ldots, A_p] = \sum_{i=1}^{k} P[B_i] \cdot P[A_1, \ldots, A_p \mid B_i] \quad \text{(theorem of total probability)}$$

$$\stackrel{NB}{=} \sum_{i=1}^{k} P[B_i] \cdot \prod_{j=1}^{p} P[A_j \mid B_i] \quad \text{(Naive Bayes Assumption)}$$

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

$$P[B_i \mid A_1, \ldots, A_p] = \frac{P[B_i] \cdot P[A_1, \ldots, A_p \mid B_i]}{P[A_1, \ldots, A_p]} \stackrel{NB,c,d}{=} \frac{P[B_i] \cdot \prod_{j=1}^{p} P[A_j \mid B_i]}{\sum_{i=1}^{k} P[B_i] \cdot \prod_{j=1}^{p} P[A_j \mid B_i]}$$

Remarks:

- A ranking of the $B_1, \ldots, B_k$ can be computed via $\underset{B \in \{B_1,\ldots,B_k\}}{\text{argmax}} P[B] \cdot \prod_{j=1}^{p} P[A_j \mid B]$.

- If both (c) completeness and (d) mutually exclusiveness of the B_i can be presumed, the total of all a-posteriori probabilities must add up to one: $\sum_{i=1}^{k} P[B_i \mid A_1, \ldots, A_p] = 1$.
  As a consequence, the rank order values of the B_i can be converted into the a-priori probabilities $P[B_i \mid A_1, \ldots, A_p]$. The normalization is obtained by dividing a rank order value by the rank order values total, i.e., $\sum_{i=1}^{k} P[B_i] \cdot \prod_{j=1}^{p} P[A_j \mid B_i]$.

Bayesian Learning: Naive Bayes

Naive Bayes: Classifier Construction Summary

Let X be a p-dimensional feature space, let Y be the set of k classes of a target concept, and let D be a set of examples of the form $(\mathbf{x}, y(\mathbf{x}))$ over $X \times Y$. Then the k classes correspond to the events $B_1, \ldots, B_k$, and the p feature values of some $\mathbf{x} \in X$ correspond to the events $A_{1=x_1}, \ldots, A_{p=x_p}$.

Bayesian Learning: Naive Bayes

Naive Bayes: Classifier Construction Summary

Let X be a p-dimensional feature space, let Y be the set of k classes of a target concept, and let D be a set of examples of the form $(\mathbf{x}, y(\mathbf{x}))$ over $X \times Y$. Then the k classes correspond to the events $B_1, \ldots, B_k$, and the p feature values of some $\mathbf{x} \in X$ correspond to the events $A_{1=x_1}, \ldots, A_{p=x_p}$.

Construction and application of a Naive Bayes classifier:

1. Estimation of the $P[B_i]$, where $B_i = y(\mathbf{x})$, $(\mathbf{x}, y(\mathbf{x})) \in D$.

2. Estimation of the $P[A_{j=x_j} \mid B_i]$, where $x_j \in \mathbf{x}$, $(\mathbf{x}, y(\mathbf{x})) \in D$, $y(\mathbf{x}) = B_i$.

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

$$B_{NB} = \underset{B \in \{B_1, \ldots, B_k\}}{\text{argmax}} \; \hat{P}(B) \cdot \prod_{\substack{x_j \in \mathbf{x} \\ j=1,\ldots,p}} \hat{P}(A_{j=x_j} \mid B)$$

Remarks:

- There are at most $p \cdot l$ different events $A_{j=x_j}$, if l is an upper bound for the size of the p feature domains.
- The probabilities, denoted as $P[\_]$, are unknown and estimated by the relative frequencies, denoted as $\hat{P}(\_)$.
- 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 a-posteriori 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”.

Bayesian Learning: Naive Bayes

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 $y(\mathbf{x})$ of feature vector $\mathbf{x} = (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})$ be unknown.

Bayesian Learning: Naive Bayes

Naive Bayes: Example (continued)

Computation of B_{NB} for **x** :

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

$$= \underset{B \in \{\textit{yes},\textit{no}\}}{\text{argmax}} \; \hat{P}(B) \cdot \hat{P}(\textit{Outlook=sunny} \mid B) \cdot \hat{P}(\textit{Temperature=cool} \mid B) \cdot \ldots$$

"$A_{j=x_j}$" denotes the event for the respective attribute-value combination in **x**. As an example, the feature vector $\mathbf{x} = (\textit{sunny}, \textit{cool}, \textit{high}, \textit{strong})$ gives rise to the following four events:

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

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

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

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

Bayesian Learning: Naive Bayes

Naive Bayes: Example (continued)

For the classification of **x** altogether $2 + 4 \cdot 2$ probabilities are to be estimated:

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

Bayesian Learning: Naive Bayes

Naive Bayes: Example (continued)

For the classification of **x** altogether $2 + 4 \cdot 2$ probabilities are 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. $\hat{P}(\textit{EnjoySport=no}) \cdot \prod_{x_j \in \mathbf{x}} \hat{P}(A_{j=x_j} \mid \textit{EnjoySport=no}) = 0.0206$

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

Naive Bayes: Example (continued)

For the classification of **x** altogether $2 + 4 \cdot 2$ probabilities are to be estimated:

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

→ Ranking:

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

→ Normalization: (subject to conditions (c) and (d))

1. $\hat{P}(\mathit{EnjoySport{=}no} \mid \mathbf{x}) = \frac{0.0206}{0.0053+0.0206} = 0.795$
2. $\hat{P}(\mathit{EnjoySport{=}yes} \mid \mathbf{x}) = \frac{0.0053}{0.0053+0.0206} = 0.205$