Bayesian Learning

[Read Ch. 6]
[Suggested exercises: 6.1, 6.2, 6.6]

• Bayes Theorem
• MAP, ML hypotheses
• MAP learners
• Minimum description length principle
• Bayes optimal classifier
• Naive Bayes learner
• Example: Learning over text data
• Bayesian belief networks
• Expectation Maximization algorithm
Two Roles for Bayesian Methods

Provides practical learning algorithms:

- Naive Bayes learning
- Bayesian belief network learning
- Combine prior knowledge (prior probabilities) with observed data
- Requires prior probabilities

Provides useful conceptual framework

- Provides “gold standard” for evaluating other learning algorithms
- Additional insight into Occam’s razor
Bayes Theorem

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]

- \( P(h) \) = prior probability of hypothesis \( h \)
- \( P(D) \) = prior probability of training data \( D \)
- \( P(h|D) \) = probability of \( h \) given \( D \)
- \( P(D|h) \) = probability of \( D \) given \( h \)
Choosing Hypotheses

\[ P(h|D) = \frac{P(D|h)P(h)}{P(D)} \]

Generally want the most probable hypothesis given the training data

**Maximum a posteriori hypothesis** \( h_{MAP} \):

\[ h_{MAP} = \arg \max_{h \in H} P(h|D) \]

\[ = \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \]

\[ = \arg \max_{h \in H} P(D|h)P(h) \]

If assume \( P(h_i) = P(h_j) \) then can further simplify, and choose the **Maximum likelihood** (ML) hypothesis

\[ h_{ML} = \arg \max_{h_i \in H} P(D|h_i) \]
Bayes Theorem

Does patient have cancer or not?

A patient takes a lab test and the result comes back positive. The test returns a correct positive result in only 98% of the cases in which the disease is actually present, and a correct negative result in only 97% of the cases in which the disease is not present. Furthermore, .008 of the entire population have this cancer.

\[
\begin{align*}
P(\text{cancer}) &= \quad P(\neg \text{cancer}) = \\
P(+) | \text{cancer} &= \quad P(-) | \text{cancer} = \\
P(+) | \neg \text{cancer} &= \quad P(-) | \neg \text{cancer} = 
\end{align*}
\]
Basic Formulas for Probabilities

- **Product Rule**: probability $P(A \land B)$ of a conjunction of two events $A$ and $B$:
  \[ P(A \land B) = P(A|B)P(B) = P(B|A)P(A) \]

- **Sum Rule**: probability of a disjunction of two events $A$ and $B$:
  \[ P(A \lor B) = P(A) + P(B) - P(A \land B) \]

- **Theorem of total probability**: if events $A_1, \ldots, A_n$ are mutually exclusive with $\sum_{i=1}^{n} P(A_i) = 1$, then
  \[ P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i) \]
Brute Force MAP Hypothesis Learner

1. For each hypothesis $h$ in $H$, calculate the posterior probability

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

2. Output the hypothesis $h_{MAP}$ with the highest posterior probability

$$h_{MAP} = \arg\max_{h \in H} P(h|D)$$
Most Probable Classification of New Instances

So far we’ve sought the most probable hypothesis given the data $D$ (i.e., $h_{MAP}$)

Given new instance $x$, what is its most probable classification?

- $h_{MAP}(x)$ is not the most probable classification!

Consider:

- Three possible hypotheses:
  \[ P(h_1|D) = .4, \quad P(h_2|D) = .3, \quad P(h_3|D) = .3 \]
- Given new instance $x$,
  \[ h_1(x) = +, \quad h_2(x) = -, \quad h_3(x) = - \]
- What’s most probable classification of $x$?
Bayes Optimal Classifier

Bayes optimal classification:

\[
\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)
\]

Example:

\[
P(h_1|D) = .4, \quad P(-|h_1) = 0, \quad P(+|h_1) = 1
\]
\[
P(h_2|D) = .3, \quad P(-|h_2) = 1, \quad P(+|h_2) = 0
\]
\[
P(h_3|D) = .3, \quad P(-|h_3) = 1, \quad P(+|h_3) = 0
\]

therefore

\[
\sum_{h_i \in H} P(+|h_i)P(h_i|D) = .4
\]
\[
\sum_{h_i \in H} P(-|h_i)P(h_i|D) = .6
\]

and

\[
\arg \max_{v_j \in V} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) = -
\]
Gibbs Classifier

Bayes optimal classifier provides best result, but can be expensive if many hypotheses.

Gibbs algorithm:

1. Choose one hypothesis at random, according to \(P(h|D)\)
2. Use this to classify new instance

Surprising fact: Assume target concepts are drawn at random from \(H\) according to priors on \(H\). Then:

\[
E[error_{Gibbs}] \leq 2E[error_{BayesOptimal}]
\]

Suppose correct, uniform prior distribution over \(H\), then

- Pick any hypothesis from VS, with uniform probability
- Its expected error no worse than twice Bayes optimal
Naive Bayes Classifier

Along with decision trees, neural networks, nearest nbr, one of the most practical learning methods.

When to use

• Moderate or large training set available
• Attributes that describe instances are conditionally independent given classification

Successful applications:

• Diagnosis
• Classifying text documents
Naive Bayes Classifier

Assume target function $f : X \rightarrow V$, where each instance $x$ described by attributes $\langle a_1, a_2 \ldots a_n \rangle$. Most probable value of $f(x)$ is:

$$v_{MAP} = \arg \max_{v_j \in V} P(v_j | a_1, a_2 \ldots a_n)$$

$$v_{MAP} = \arg \max_{v_j \in V} \frac{P(a_1, a_2 \ldots a_n | v_j) P(v_j)}{P(a_1, a_2 \ldots a_n)} = \arg \max_{v_j \in V} P(a_1, a_2 \ldots a_n | v_j) P(v_j)$$

Naive Bayes assumption:

$$P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j)$$

which gives

**Naive Bayes classifier:** $v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_i P(a_i | v_j)$
Naive Bayes Algorithm

Naive_Bayes_Learn(examples)
For each target value $v_j$
    $\hat{P}(v_j) \leftarrow$ estimate $P(v_j)$
For each attribute value $a_i$ of each attribute $a$
    $\hat{P}(a_i|v_j) \leftarrow$ estimate $P(a_i|v_j)$

Classify_New_Instance($x$)

$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_{a_i \in x} \hat{P}(a_i|v_j)$
Naive Bayes: Example

Consider *PlayTennis* again, and new instance

\[ \langle \text{Outlk} = \text{sun}, \text{Temp} = \text{cool}, \text{Humid} = \text{high}, \text{Wind} = \text{strong} \rangle \]

Want to compute:

\[
v_{NB} = \arg\max_{v_j \in V} P(v_j) \prod_{i} P(a_i | v_j)
\]

\[
P\left(y\right) P\left(\text{sun} | y\right) P\left(\text{cool} | y\right) P\left(\text{high} | y\right) P\left(\text{strong} | y\right) = .005
\]

\[
P\left(n\right) P\left(\text{sun} | n\right) P\left(\text{cool} | n\right) P\left(\text{high} | n\right) P\left(\text{strong} | n\right) = .021
\]

\[
\rightarrow v_{NB} = n
\]
Naive Bayes: Subtleties

1. Conditional independence assumption is often violated

\[ P(a_1, a_2 \ldots a_n | v_j) = \prod_i P(a_i | v_j) \]

• ...but it works surprisingly well anyway. Note don’t need estimated posteriors \( \hat{P}(v_j|x) \) to be correct; need only that

\[ \arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i | v_j) = \arg\max_{v_j \in V} P(v_j) P(a_1 \ldots , a_n | v_j) \]

• see [Domingos & Pazzani, 1996] for analysis
• Naive Bayes posteriors often unrealistically close to 1 or 0
Naive Bayes: Subtleties

2. what if none of the training instances with target value \( v_j \) have attribute value \( a_i \)? Then

\[
\hat{P}(a_i|v_j) = 0, \text{ and...}
\]

\[
\hat{P}(v_j) \prod_i \hat{P}(a_i|v_j) = 0
\]

Typical solution is Bayesian estimate for \( \hat{P}(a_i|v_j) \)

\[
\hat{P}(a_i|v_j) \leftarrow \frac{n_c + mp}{n + m}
\]

where

- \( n \) is number of training examples for which \( v = v_j \),
- \( n_c \) number of examples for which \( v = v_j \) and \( a = a_i \),
- \( p \) is prior estimate for \( \hat{P}(a_i|v_j) \)
- \( m \) is weight given to prior (i.e. number of “virtual” examples)
Learning to Classify Text

Why?

• Learn which news articles are of interest
• Learn to classify web pages by topic

Naive Bayes is among most effective algorithms

What attributes shall we use to represent text documents??
Learning to Classify Text

Target concept $Interesting?: Document \rightarrow \{+, -\}$

1. Represent each document by vector of words
   - one attribute per word position in document
2. Learning: Use training examples to estimate
   - $P(+)$
   - $P(-)$
   - $P(doc|+)$
   - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|v_j) = \prod_{i=1}^{length(doc)} P(a_i = w_k|v_j)$$

where $P(a_i = w_k|v_j)$ is probability that word in position $i$ is $w_k$, given $v_j$

one more assumption:

$$P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$$
**Learn naive Bayes text** (*Examples, V*)

1. collect all words and other tokens that occur in *Examples*

* Vocabulary $\leftarrow$ all distinct words and other tokens in *Examples*

2. calculate the required $P(v_j)$ and $P(w_k|v_j)$ probability terms

* For each target value $v_j$ in $V$ do
  
  $- docs_j \leftarrow$ subset of *Examples* for which the target value is $v_j$

  $- P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$

  $- Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$

  $- n \leftarrow$ total number of words in $Text_j$ (counting duplicate words multiple times)

  $- for$ each word $w_k$ in Vocabulary

  $* n_k \leftarrow$ number of times word $w_k$ occurs in $Text_j$

  $* P(w_k|v_j) \leftarrow \frac{n_k + 1}{n + |Vocabulary|}$
\textbf{CLASSIFY\_NAIVE\_BAYES\_TEXT}(Doc)

\begin{itemize}
  \item \textit{positions} \leftarrow \text{all word positions in } Doc \text{ that contain tokens found in } Vocabulary
  \item Return $v_{NB}$, where
    
    \[
v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i \in \text{positions}} P(a_i | v_j)
    \]
\end{itemize}
Twenty NewsGroups

Given 1000 training documents from each group
Learn to classify new documents according to
which newsgroup it came from

comp.graphics    misc.forsale
comp.os.ms-windows.misc    rec.autos
comp.sys.ibm.pc.hardware    rec.motorcycles
comp.sys.mac.hardware    rec.sport.baseball
comp.windows.x    rec.sport.hockey

alt.atheism
soc.religion.christian
talk.religion.misc
talk.politics.mideast
talk.politics.misc
talk.politics.guns

sci.space
sci.crypt
sci.electronics
sci.med

Naive Bayes: 89% classification accuracy
I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he’s clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he’s only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided
Learning Curve for 20 Newsgroups

Accuracy vs. Training set size (1/3 withheld for test)