MLE’s, Bayesian Classifiers and Naïve Bayes

Required reading:

- Mitchell draft chapter, sections 1 and 2. (available on class website)

Machine Learning 10-601

Tom M. Mitchell
Machine Learning Department
Carnegie Mellon University

January 30, 2008
Naïve Bayes in a Nutshell

Bayes rule:

\[
P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) P(X_1 \ldots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \ldots X_n | Y = y_j)}
\]

Assuming conditional independence among \( X_i \)'s:

\[
P(Y = y_k | X_1 \ldots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}
\]

So, classification rule for \( X^{\text{new}} = < X_1, \ldots, X_n > \) is:

\[
Y^{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i^{\text{new}} | Y = y_k)
\]
Naïve Bayes Algorithm – discrete $X_i$

• Train Naïve Bayes (examples)
  for each* value $y_k$
    estimate $\pi_k \equiv P(Y = y_k)$
  for each* value $x_{ij}$ of each attribute $X_i$
    estimate $\theta_{ijk} \equiv P(X_i = x_{ij} | Y = y_k)$

• Classify ($X_{\text{new}}$)
  $Y_{\text{new}} \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_{i\text{new}} | Y = y_k)$
  $Y_{\text{new}} \leftarrow \arg \max_{y_k} \pi_k \prod_i \theta_{ijk}$

* probabilities must sum to 1, so need estimate only n-1 parameters...
Estimating Parameters: $Y, X_i$ discrete-valued

Maximum likelihood estimates (MLE’s):

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij}|Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}$$

Number of items in set $D$ for which $Y=y_k$
Example: Live in Sq Hill? \( P(S|G,D,M) \)

- \( S=1 \) iff live in Squirrel Hill
- \( G=1 \) iff shop at Giant Eagle
- \( D=1 \) iff Drive to CMU
- \( M=1 \) iff Dave Matthews fan

\[
P(S=1) = \frac{3}{43}
\]

\[
P(S=0) = 1 - P(S=1)
\]

\[
P(G=1 | S=1) = \frac{6}{4}
\]

\[
P(G=1 | S=0) = \frac{11}{40}
\]

\[
P(D=1 | S=1) = \frac{2}{8}
\]

\[
P(D=1 | S=0) = \frac{3}{40}
\]

\[
P(M=1 | S=1) = \frac{1}{8}
\]

\[
P(M=1 | S=0) = \frac{6}{40}
\]

\[
P(S=1 | G=1, D=1, M=0) = \frac{4}{8}
\]

\[
P(S=1 | G=1, D=1, M=1) = \frac{8}{43}
\]

\[
P(S=1 | G=0, D=1, M=0) = \frac{6}{8}
\]

\[
P(S=1 | G=0, D=1, M=1) = \frac{3}{40}
\]

\[
P(S=1 | G=0, D=0, M=0) = \frac{3}{40}
\]

\[
P(S=1 | G=0, D=0, M=1) = \frac{34}{40}
\]

\[
Q_2 = \frac{146}{40} = P(S=0) \cdot P(G=1 | S=0) \cdot P(D=1 | S=0) \cdot P(M=0 | S=0)
\]

\[
\frac{40}{43} \cdot \frac{11}{40} \cdot \frac{3}{40} \cdot \frac{34}{40}
\]
Example: Live in Sq Hill? P(S|G,D,M)

- S=1 iff live in Squirrel Hill
- G=1 iff shop at Giant Eagle
- D=1 iff Drive to CMU
- M=1 iff Dave Matthews fan
Naïve Bayes: Subtlety #1

If unlucky, our MLE estimate for $P(X_i \mid Y)$ may be zero. (e.g., $X_{373} =$ Birthday Is January30)

\[ \hat{P}(X=1 \mid S=1) = 0 = \hat{P}(X=1 \mid S=0) \]

• Why worry about just one parameter out of many?

• What can be done to avoid this?
Estimating Parameters: $Y, X_i$, discrete-valued

Maximum likelihood estimates:

$$\hat{\pi}_k = \hat{P}(Y = y_k) = \frac{\#D\{Y = y_k\}}{|D|}$$

$$\hat{\theta}_{ijk} = \hat{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\}}{\#D\{Y = y_k\}}$$

MAP estimates (Dirichlet priors):

$$\tilde{\pi}_k = \tilde{P}(Y = y_k) = \frac{\#D\{Y = y_k\} + lR}{|D| + lR}$$

$$\tilde{\theta}_{ijk} = \tilde{P}(X_i = x_{ij} | Y = y_k) = \frac{\#D\{X_i = x_{ij} \land Y = y_k\} + lM}{\#D\{Y = y_k\} + lM}$$

Only difference: “imaginary” examples
Naïve Bayes: Subtlety #2

Often the $X_i$ are not really conditionally independent

- We use Naïve Bayes in many cases anyway, and it often works pretty well
  - often the right classification, even when not the right probability (see [Domingos&Pazzani, 1996])

- What is effect on estimated $P(Y|X)$?
  - Special case: what if we add two copies: $X_i = X_k$

$$Q_1 = P(Y)P(X_i | Y)P(X_2 | Y)P(X_3 | Y)P(X_4 | Y)$$

$$Q_1 + Q_2$$
Learning to classify text documents

- Classify which emails are spam
- Classify which emails are meeting invites
- Classify which web pages are student home pages

How shall we represent text documents for Naïve Bayes?
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 to Classify Text

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

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

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

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

one more assumption:

\[
P(a_i = w_k | v_j) = P(a_m = w_k | v_j), \forall i, m
\]
Baseline: Bag of Words Approach

Our energy exploration, production, and distribution operations span the globe, with activities in more than 100 countries.

At TOTAL, we draw our greatest strength from our fast-growing oil and gas reserves. Our strategic emphasis on natural gas provides a strong position in a rapidly expanding market.

Our expanding refining and marketing operations in Asia and the Mediterranean Rim complement already solid positions in Europe, Africa, and the U.S.

Our growing specialty chemicals sector adds balance and profit to the core energy business.

<table>
<thead>
<tr>
<th>Word</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>aardvark</td>
<td>0</td>
</tr>
<tr>
<td>about</td>
<td>2</td>
</tr>
<tr>
<td>all</td>
<td>2</td>
</tr>
<tr>
<td>Africa</td>
<td>1</td>
</tr>
<tr>
<td>apple</td>
<td>0</td>
</tr>
<tr>
<td>anxious</td>
<td>0</td>
</tr>
<tr>
<td>gas</td>
<td>1</td>
</tr>
<tr>
<td>oil</td>
<td>1</td>
</tr>
<tr>
<td>Zaire</td>
<td>0</td>
</tr>
</tbody>
</table>
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  sci.space
soc.religion.christian  sci.crypt
talk.religion.misc  sci.electronics
talk.politics.mideast  sci.med
talk.politics.misc
talk.politics.guns

Naive Bayes: 89% classification accuracy
\textsc{Learn naive Bayes text}(Examples, V)

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

\textbullet\ Vocabulary \leftarrow \text{all distinct words and other tokens in Examples}

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

\textbullet\ For each target value $v_j$ in $V$ do

\hspace{1cm} - docs$_j$ \leftarrow \text{subset of Examples for which the target value is } v_j

\hspace{1cm} - P(v_j) \leftarrow \frac{|\text{docs}_j|}{|\text{Examples}|}

\hspace{1cm} - Text$_j$ \leftarrow \text{a single document created by concatenating all members of docs}_j

\hspace{1cm} - n \leftarrow \text{total number of words in Text}_j \text{ (counting duplicate words multiple times)}

\hspace{1cm} - \text{for each word } w_k \text{ in Vocabulary}

\hspace{1.5cm} * n_{k,j} \leftarrow \text{number of times word } w_k \text{ occurs in Text}_j

\hspace{1.5cm} * P(w_k|v_j) \leftarrow \frac{n_{k,j}+1}{n+|\text{Vocabulary}|}
CLASSIFY NAIVE BAYES TEXT(Doc)

• \textit{positions} \leftarrow \text{all word positions in } Doc \text{ that contain tokens found in } Vocabular y

• Return \( v_{NB} \), where

\[
v_{NB} = \arg \max_{v_j \in V} P(v_j) \prod_{i \in \text{positions}} P(a_i|v_j)
\]
Learning Curve for 20 Newsgroups

Accuracy vs. Training set size (1/3 withheld for test)
What you should know:

• Training and using classifiers based on Bayes rule

• Conditional independence
  – What it is
  – Why it’s important

• Naïve Bayes
  – What it is
  – Why we use it so much
  – Training using MLE, MAP estimates
  – Discrete variables (Bernoulli) and continuous (Gaussian)
Questions:

• Can you use Naïve Bayes for a combination of discrete and real-valued $X_i$?

• How can we easily model just 2 of n attributes as dependent?

• What does the decision surface of a Naïve Bayes classifier look like?