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 $X^{new} = < X_1, \ldots, X_n >$ is:

$$Y^{new} \leftarrow \arg\max_{y_k} \; P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

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^{new})

$$Y^{new} \leftarrow \arg\max_{y_k} \ P(Y = y_k) \prod_i P(X_i^{new} | Y = y_k)$$

$$Y^{new} \leftarrow \arg\max_{y_k} \ \pi_k \prod_i \theta_{ijk}$$

* probabilities must sum to 1, so need estimate only n-1 parameters...

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates (MLE's):

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

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

Number of items in set D for which $Y=y_k$

Example: Live in Sq Hill? P(S|G,D,M)

$P(S=1 \mid test\ 1) = \frac{Q1}{Q1+Q2}$

- 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

$P(S=1) = 8/48$ $\quad P(S=0) = 1 - P(S=1) = 40/48$

$P(G=1 \mid S=1)\ 6/8 \quad P(G=0 \mid S=1) = 1 - \ldots$

$P(G=1 \mid S=0)\ 11/40$

$P(D=1 \mid S=1)\ 2/8$

$P(D=1 \mid S=0)\ 3/40$

$P(M=1 \mid S=1)\ 1/8 \quad 1 - 1/8 = 7/8$

$P(M=1 \mid S=0)\ 6/40$

$\boxed{.0273 = Q1}$

$P(S=1 \mid G=1, D=1, M=0)$

$\propto \{P(S=1)\, P(G=1 \mid S=1)\, P(D=1 \mid S=1)\, P(M=0 \mid S=1)\}$ $\quad 8/48 \quad 6/8 \quad 2/8 \quad 7/8$

test	G	D	M	P(S=1|test)	S true
1	1	1	0	2/3	0
2	0	0	0	.02	0
3	1	1	1	.60	1

$Q2 = .0146 = P(S=0) \cdot P(G=1 \mid S=0) \cdot P(D=1 \mid S=0)\, P(M=0 \mid S=0)$

$40/48 \cdot 11/40 \cdot 3/40 \cdot 34/40$

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)

$\hat{P}(x=1 \mid S=1) = 0 = \hat{P}(x=0 \mid S=0)$

- Why worry about just one parameter out of many?

- What can be done to avoid this?

Estimating Parameters: Y, X_i discrete-valued

maximize P(data | θ)

Maximum likelihood estimates:

\# of examples in dataset D that satisfy Y=y_k

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

40

48

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

max P(θ | data) = P(data | θ) P(θ)

MAP estimates (Dirichlet priors):

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

Only difference: "imaginary" examples

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

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: $X_i = X_k$

$$\frac{Q_1 = P(Y)\,P(X_1|Y)\,P(X_2|Y)\,P(\overset{.7}{X_3}|Y)\,P(\overset{.7}{X_4}|Y)}{Q1 + Q2}$$

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?

Article from rec.sport.hockey

```
Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.e
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinic
Date: 5 Apr 93 09:53:39 GMT
```

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

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:
$P(a_i = w_k|v_j) = P(a_m = w_k|v_j), \forall i, m$

Baseline: Bag of Words Approach

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	1
...	
oil	1
…	
Zaire	0

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

$\text{LEARN\_NAIVE\_BAYES\_TEXT}(Examples, V)$

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 $Text_j$
    * $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

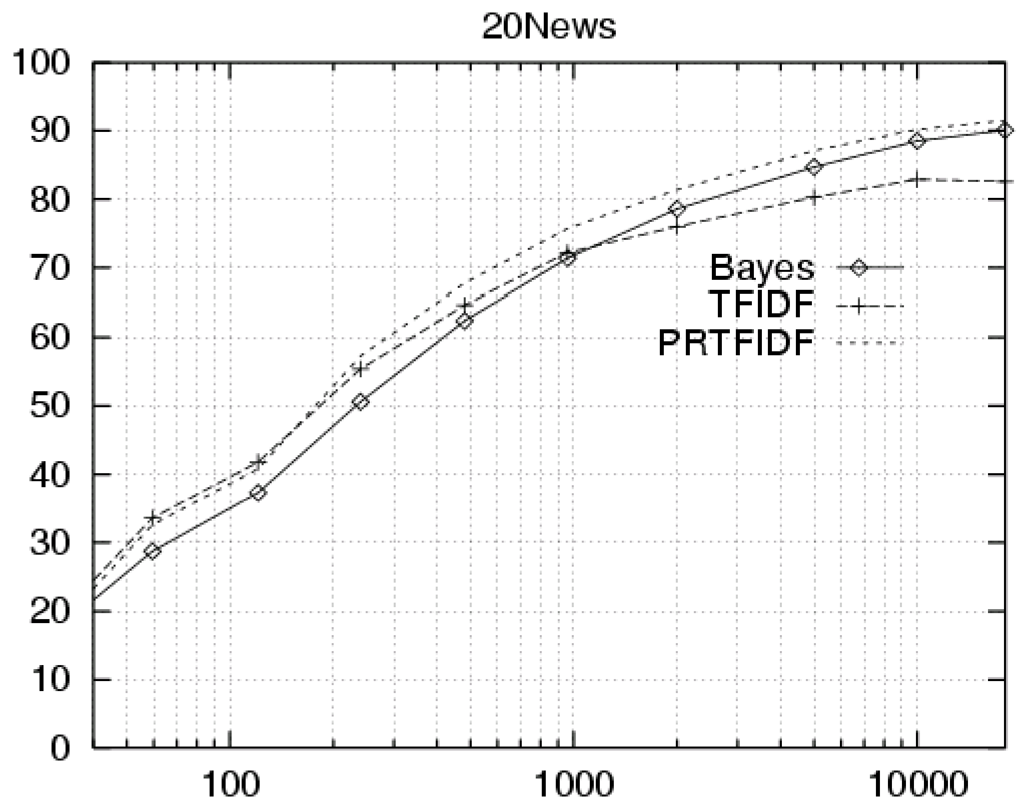
For code and data, see
www.cs.cmu.edu/~tom/mlbook.html
click on "Software and Data"

CLASSIFY_NAIVE_BAYES_TEXT(Doc)

- $positions \leftarrow$ all word positions in Doc that contain tokens found in $Vocabulary$
- Return v_{NB}, where

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_{i \in positions} P(a_i | v_j)$$

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)

$P(X_i | Y)$

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?