Naïve Bayes and Logistic Regression

Required reading:

- Mitchell draft chapter (see course website)

Recommended reading:

- Mitchell, 6.10 (text learning example)
- Bishop, Chapter 3.1.3, 3.1.4
- Ng and Jordan paper

Machine Learning 10-701

Tom M. Mitchell
Center for Automated Learning and Discovery
Carnegie Mellon University

September 27, 2005

Naïve Bayes and Logistic Regression

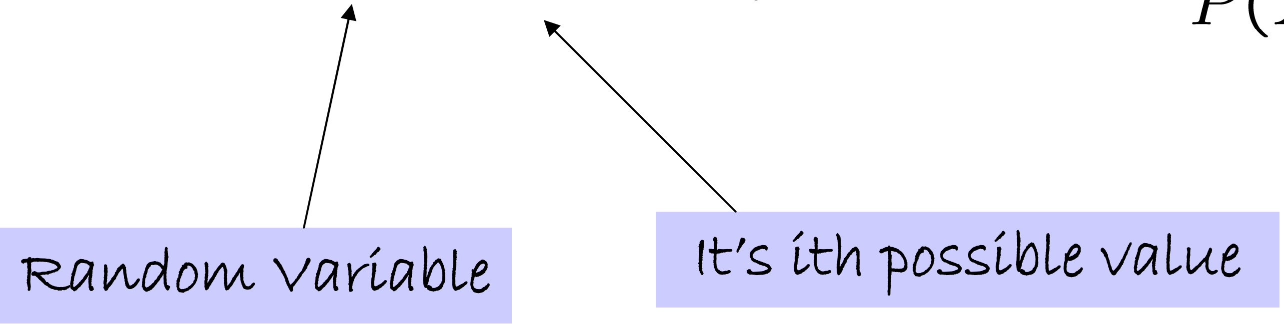
- Design learning algorithms based on our understanding of probability
- Two of the most widely used
- Interesting relationship between these two
- Generative and Discriminative classifiers

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i,j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{P(X = x_j)}$$



Random variable

It's ith possible value

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

$$(\forall i,j)P(Y=y_i|X=x_j) = \frac{P(X=x_j|Y=y_i)P(Y=y_i)}{P(X=x_j)}$$

Equivalently:

$$(\forall i,j)P(Y=y_i|X=x_j) = \frac{P(X=x_j|Y=y_i)P(Y=y_i)}{\sum_k P(X=x_j|Y=y_k)P(Y=y_k)}$$

Bayes Rule

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Which is shorthand for:

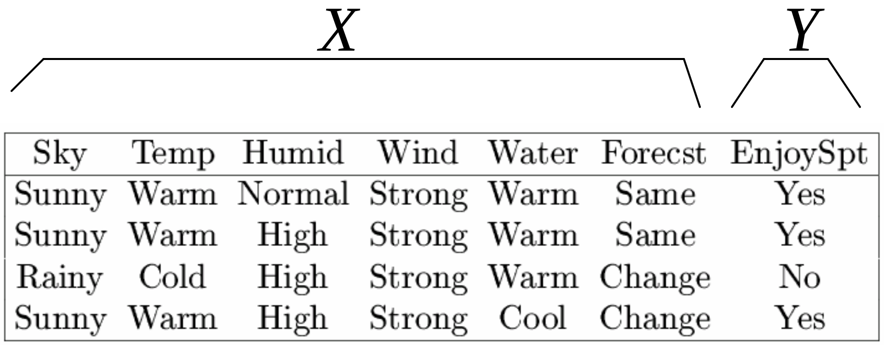
$$(\forall i,j) P(Y=y_i|X=x_j) = \frac{P(X=x_j|Y=y_i)P(Y=y_i)}{P(X=x_j)}$$

Common abbreviation:

$$(\forall i,j)\ \ P(y_i|x_j) = \frac{P(x_j|y_i)P(y_i)}{P(x_j)}$$

Bayes Classifier

Training data:



Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

Learning = estimating P(X|Y), P(Y)

Classification = using Bayes rule to calculate P(Y | X^{new})

Bayes Classifier

Training data:

X						Y
Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

How shall we represent P(X|Y), P(Y)?

How many parameters must we estimate?

Bayes Classifier

Training data:

$\overbrace{\qquad\qquad\qquad\qquad}^{X}$ $\overbrace{\qquad}^{Y}$

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

How shal … esent P(X|Y), P(Y)?

How … arameters must we estimate?

Full joint $P(X_1 \ldots X_n | Y)$
usually impractical!

Naïve Bayes

Naïve Bayes assumes

$X=\langle X_1, \ldots, X_n \rangle$, Y discrete-valued

$$P(X_1 \ldots X_n|Y) = \prod_i P(X_i|Y)$$

i.e., that X_i and X_j are conditionally independent given Y, for all $i \neq j$

Conditional Independence

Definition: X is conditionally independent of Y given Z, if the probability distribution governing X is independent of the value of Y, given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write

$$P(X|Y, Z) = P(X|Z)$$

E.g.,

$$P(Thunder|Rain, Lightning) = P(Thunder|Lightning)$$

Naïve Bayes uses assumption that the X_i are conditionally independent, given Y

then:

$$P(X_1, X_2|Y) = P(X_1|X_2, Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

$$P(X_1 ... X_n|Y) = \prod_i P(X_i|Y)$$

How many parameters needed now for P(X|Y)? P(Y)?

$$\theta_{ij} \equiv P(X = x_i|Y = y_j) \qquad \pi_j \equiv P(Y = y_j)$$

Naïve Bayes classification

Bayes rule:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

Assuming conditional independence:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

So, classification rule for $X^{new} = \langle X_1, \ldots, X_n \rangle$ is:

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

Naïve Bayes Algorithm

- 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})

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

* parameters must sum to 1

Estimating Parameters: Y, X_i discrete-valued

Maximum likelihood estimates:

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

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

MAP estimates (uniform Dirichlet priors):

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

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

Naive Bayes: Example

Consider *PlayTennis* again, and new instance

$\langle Outlk = sun, Temp = cool, Humid = high, Wind = strc$

Want to compute:

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

$P(y)\ P(sun|y)\ P(cool|y)\ P(high|y)\ P(strong|y) = .005$

$P(n)\ P(sun|n)\ P(cool|n)\ P(high|n)\ P(strong|n) = .021$

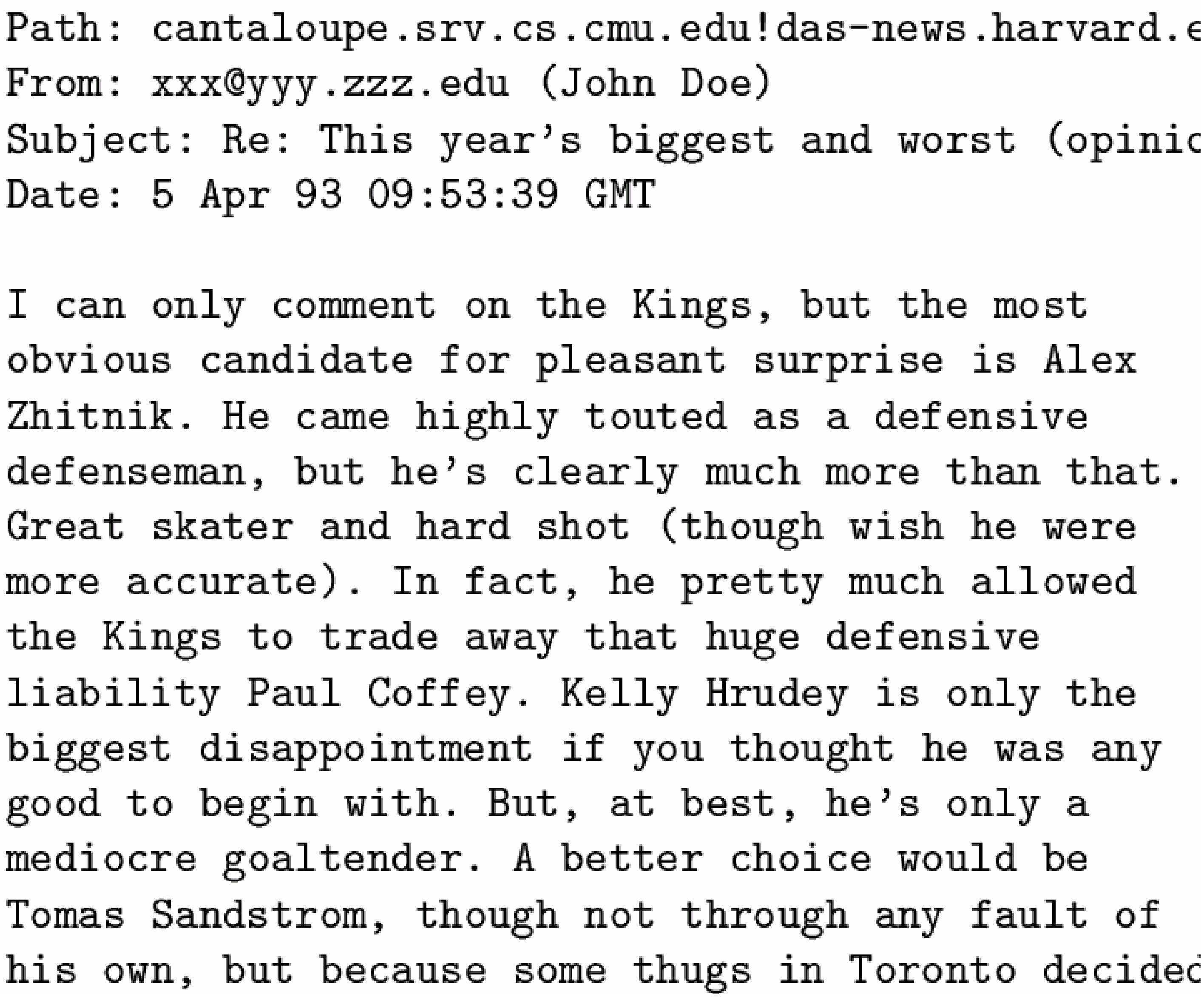
$$\rightarrow Y^{new} = n$$

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.
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opini
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 decide

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

all about the **company**

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.

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

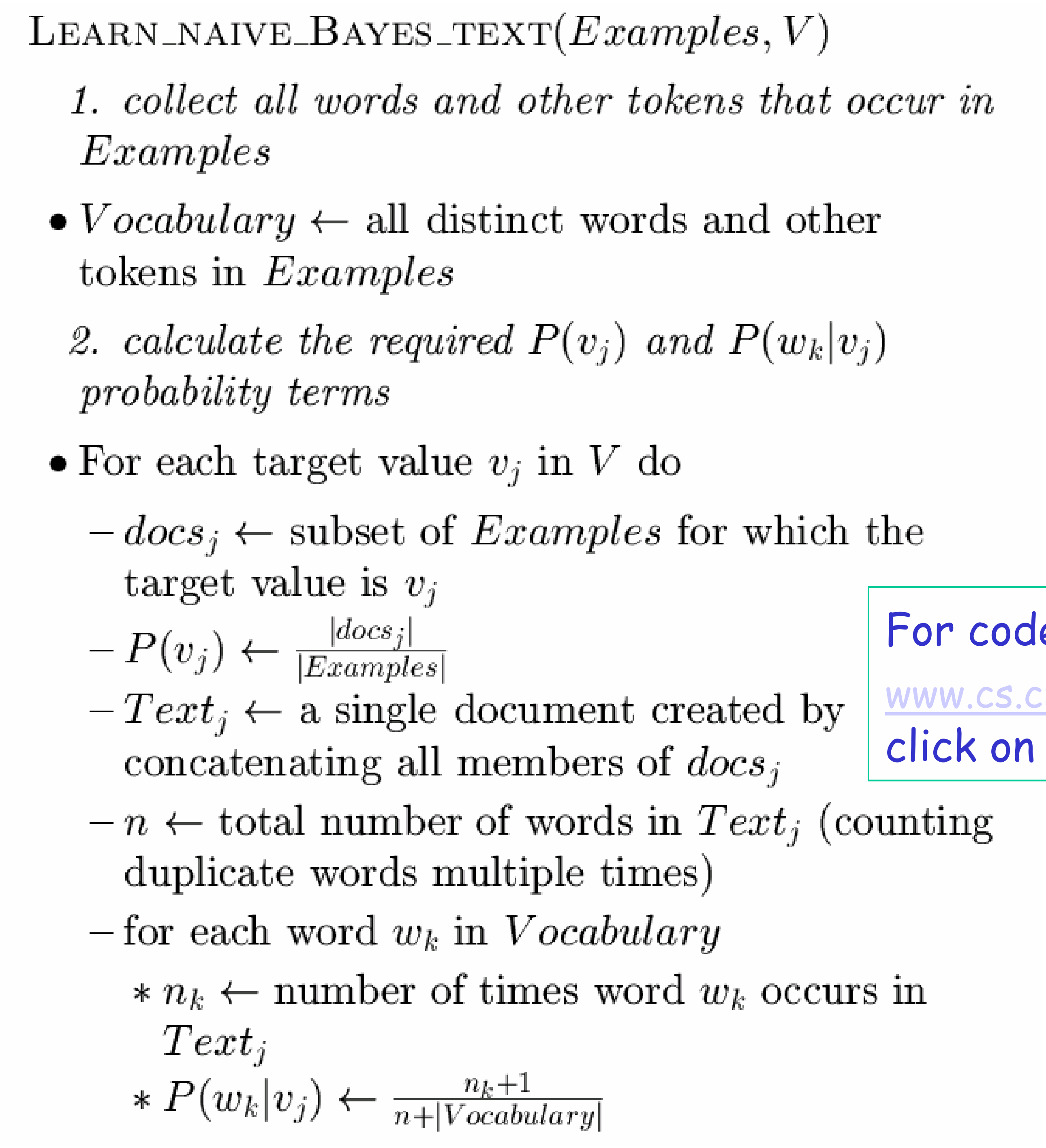
comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x

misc.forsale
rec.autos
rec.motorcycles
rec.sport.baseball
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



```
$\text{LEARN\_NAIVE\_BAYES\_TEXT}(Examples, V)$
  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 $Text_j$
      * $P(w_k|v_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$
```

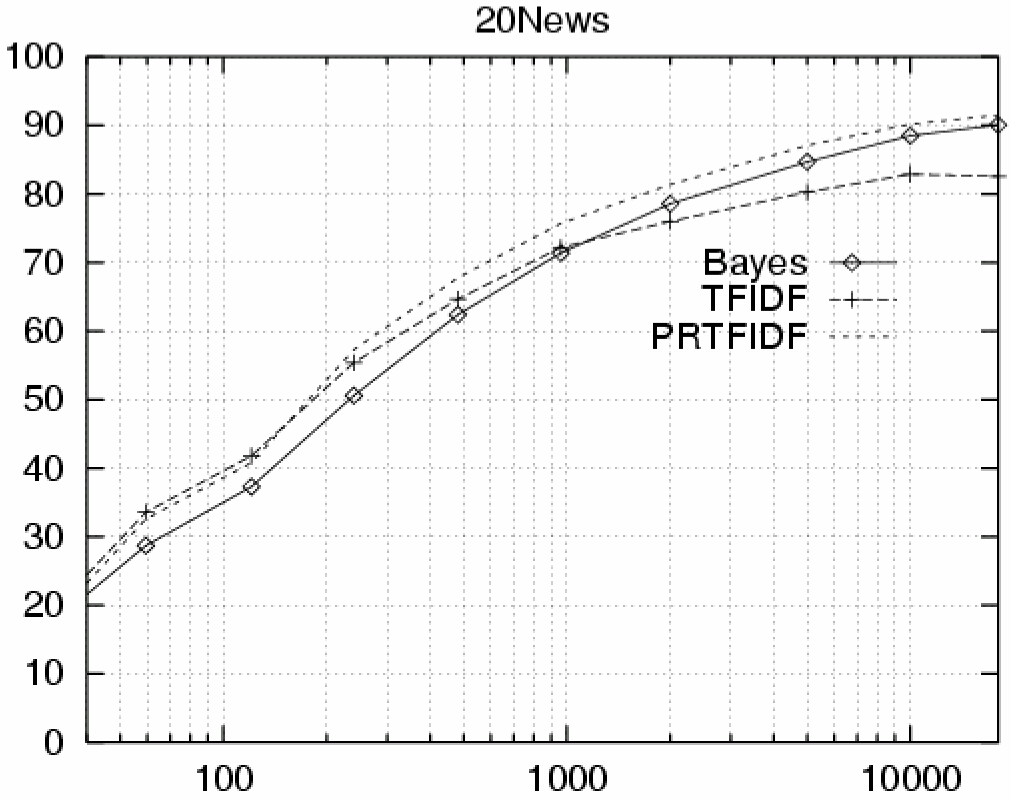
For code, see
www.cs.cmu.edu/~tom/mlbook.html
click on “Software and Data”

CLASSIFY_NAIVE_BAYES_TEXT(*Doc*)

- *positions* ← all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return v_{NB}, where

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

Learning Curve for 20 Newsgroups



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