UVA CS 6316 – Fall 2015 Graduate: Machine Learning

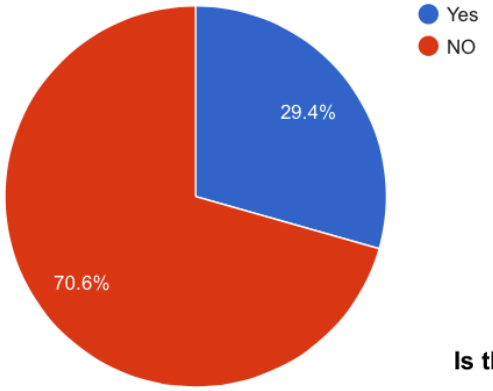
Lecture 13: Naïve Bayes Classifier (Cont.)

Dr. Yanjun Qi

University of Virginia

Department of Computer Science

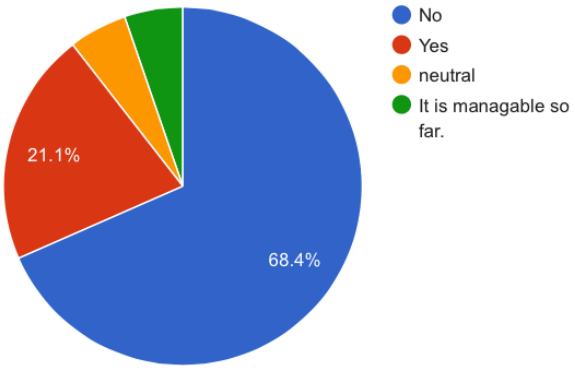
Is the teaching pace too slow ?



- HW3 posted in collab
- HW4+Solution posted in collab

Responses from Survey 2

Is the course content too difficult ?



Where are we ? ➔
Five major sections of this course

- ❑ Regression (supervised)
- ❑ Classification (supervised)
- ❑ Unsupervised models
- ❑ Learning theory
- ❑ Graphical models

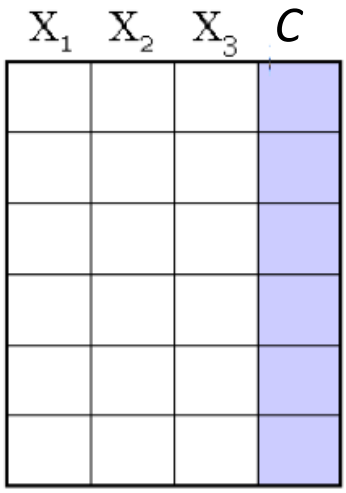
Where are we ? ➔
Three major sections for classification

- We can divide the large variety of classification approaches into roughly three major types

1. Discriminative
   - directly estimate a decision rule/boundary
   - e.g., support vector machine, decision tree

2. Generative:
   - build a generative statistical model
   - e.g., naïve bayes classifier, Bayesian networks

3. Instance based classifiers
   - Use observation directly (no models)
   - e.g. K nearest neighbors



A Dataset for classification

$$f : X \longrightarrow C$$

Output as Discrete Class Label
$C_1, C_2, \ldots, C_L$

$P(C \mid \mathbf{X})$

- **Data**/*points/instances/examples/samples/records*: [rows]
- **Features**/*attributes/dimensions/independent variables/covariates/ predictors/regressors*: [columns, except the last]
- **Target**/*outcome/response/label/dependent variable*: special column to be predicted [last column]

Bayes classifier

- Treat each attribute and class label as random variables.
- Given a sample **x** with attributes ($x_1, x_2, \ldots, x_p$):
  - Goal is to predict class *C*.
  - Specifically, we want to find the value of C_i that maximizes $p(C_i \mid x_1, x_2, \ldots, x_p)$.
- Generative Bayes classification

$$P(C \mid \mathbf{X}) \propto P(\mathbf{X} \mid C)P(C) = P(X_1, \cdots, X_p \mid C)P(C)$$

Difficulty: learning the joint probability $P(X_1, \cdots, X_p \mid C)$

Naïve Bayes Classifier

Difficulty: learning the joint probability $P(X_1,\cdots,X_p \mid C)$

- Naïve Bayes classification
  - Assumption that all input attributes are conditionally independent!

$$P(X_1,X_2,\cdots,X_p \mid C) = P(X_1 \mid X_2,\cdots,X_p,C)P(X_2,\cdots,X_p \mid C)$$
$$= P(X_1 \mid C)P(X_2,\cdots,X_p \mid C)$$
$$= P(X_1 \mid C)P(X_2 \mid C)\cdots P(X_p \mid C)$$

  - MAP classification rule: for $\mathbf{x} = (x_1,x_2,\cdots,x_p)$

$$[P(x_1 \mid c^*)\cdots P(x_p \mid c^*)]P(c^*) > [P(x_1 \mid c)\cdots P(x_p \mid c)]P(c),$$
$$c \neq c^*,\ c = c_1,\cdots,c_L$$

Adapt from Prof. Ke Chen NB slides

Naïve Bayes Assumption

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.
- $P(x_1,x_2,\ldots,x_p \mid c_j)$
  - $O(|X|^p \cdot |C|)$ parameters
  - Could only be estimated if a very, very large number of training examples was available

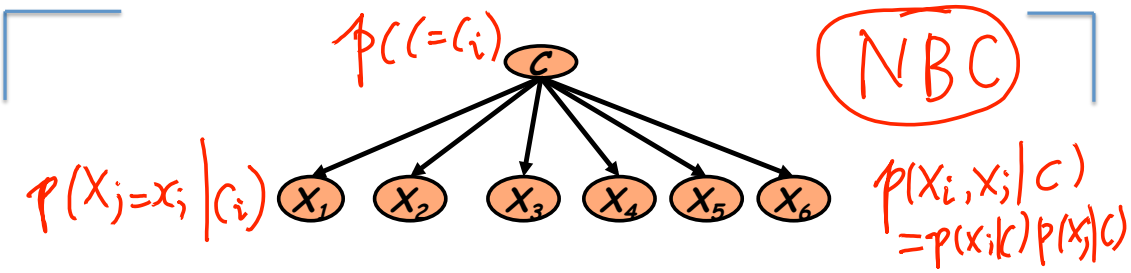
If no naïve assumption

Naïve Bayes Conditional Independence Assumption:

- Assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i \mid c_j)$.

Adapt From Manning' textCat tutorial

Learning the Model



- maximum likelihood estimates (explain later)
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{N(C = c_j)}{N}$$

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j)}{N(C = c_j)}$$

Smoothing to Avoid Overfitting

Why necessary ??

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

# of values of feature X_i

To make sum_i (P(xi|Cj)=1

$|X_i| = k$

Adapt From Manning' textCat tutorial

Smoothing to Avoid Overfitting

$$\hat{P}(x_i \mid c_j) = \frac{N(X_i = x_i, C = c_j) + 1}{N(C = c_j) + k}$$

# of values of X_i

- Somewhat more subtle version

overall fraction in data where $X_i = x_{i,k}$

$$\hat{P}(x_{i,k} \mid c_j) = \frac{N(X_i = x_{i,k}, C = c_j) + mp_{i,k}}{N(C = c_j) + m}$$

extent of "smoothing"

Today : Naïve Bayes Classifier

- ✓ Why Bayes Classification – MAP Rule?
  - Review: Mean & Variance
  - Empirical Prediction Error, 0-1 Loss function for Bayes Classifier
- ✓ Naïve Bayes Classifier for Text document categorization
  - ✓ Bag of words representation
  - ✓ Multinomial vs. multivariate Bernoulli

Bayes Classifiers – MAP Rule

Task: Classify a new instance X based on a tuple of attribute values $X = \langle X_1, X_2, \ldots, X_p \rangle$ into one of the classes

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_p)$$

WHY ?

$$= \underset{c_j \in C}{\operatorname{argmax}} \frac{P(x_1, x_2, \ldots, x_p \mid c_j) P(c_j)}{P(x_1, x_2, \ldots, x_p)}$$

$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_p \mid c_j) P(c_j)$$

MAP = Maximum Aposteriori Probability

Review: Mean and Variance of RV

- Mean (Expectation): $\mu = E(X)$
  - Discrete RVs: $E(X) = \sum_{v_i} v_i P(X = v_i)$

$$E(g(X)) = \sum_{v_i} g(v_i) P(X = v_i)$$

  - Continuous RVs: $E(X) = \int_{-\infty}^{+\infty} x f(x) dx$

$$E(g(X)) = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

Review: Mean and Variance of RV

- Variance: $Var(X) = E((X - \mu)^2)$
  - Discrete RVs: $V(X) = \sum_{v_i} (v_i - \mu)^2 P(X = v_i)$
  - Continuous RVs: $V(X) = \int_{-\infty}^{+\infty} (x - \mu)^2 f(x) dx$
- Covariance: $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$

$$Cov(X,Y) = E((X - \mu_x)(Y - \mu_y)) = E(XY) - \mu_x \mu_y$$

Review: Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
  - For discrete RV: Probability mass function (pmf): $P(X = x_i)$
- A pdf is any function $f(x)$ that describes the probability density in terms of the input variable x.

Review: Probability of Continuous RV

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

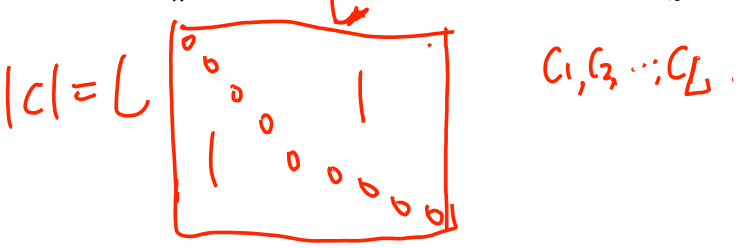
$$P(0 \leq X \leq 1) = \int_0^1 f(x)dx$$

Today : Naïve Bayes Classifier

- ✓ Why Bayes Classification – MAP Rule?
  - Review: Mean & Variance
  - Empirical Prediction Error, 0-1 Loss function for Bayes Classifier
- ✓ Naïve Bayes Classifier for Text document categorization
  - ✓ Bag of words representation
  - ✓ Multinomial vs. multivariate Bernoulli

0-1 LOSS for Classification

- Procedure for categorical output variable C
- Frequently, 0-1 loss function used: $L(k, \ell)$
- $L(k, \ell)$ is the price paid for misclassifying an element from class C_k as belonging to class C_ℓ
  - ➔ *L*L matrix*



Expected prediction error (EPE)

- Expected prediction error (EPE), with expectation taken w.r.t. the joint distribution Pr(C,X)
  - $\Pr(C,X)=\Pr(C \mid X)\Pr(X)$

$$\mathrm{EPE}(f)=E_{X,C}(L(C,f(X)))=\mathrm{E}_X\sum_{k=1}^{L}L[C_k,f(X)]\Pr(C_k \mid X)$$

Consider sample population distribution

$$EPE(f) = E_{X,C}(L(C, f(X)))$$

$$= E_X E_{C|X}[L(C, f(X)) | X]$$

Discrete RV's Expectation

$$= E_X \sum_{k=1}^{L} L[C_k, f(X)] \Pr(C_k | X)$$

Pointwise minization

$$\Rightarrow \hat{f}(X=x) = \underset{f(x) \in C}{\operatorname{argmin}} \sum_{k=1}^{L} L[C_k, f(x)] \Pr(C_k | X=x)$$

$$\Rightarrow \hat{f}(x) = \underset{C_k \in C}{\operatorname{argmax}} \Pr(C_k | X=x)$$

Expected prediction error (EPE)

$$\text{EPE}(f) = E_{X,C}(L(C, f(X))) = \text{E}_X \sum_{k=1}^{K} L[C_k, f(X)] \Pr(C_k | X)$$

Consider sample population distribution

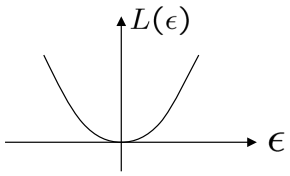
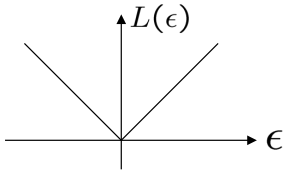
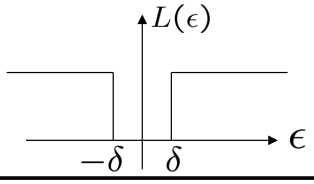
- Pointwise minimization suffices
- ➔ simply

$$\hat{f}(X) = \underset{g \in C}{\operatorname{argmin}} \sum_{k=1}^{K} L(C_k, g) \Pr(C_k | X = x)$$

$$\hat{f}(X) = C_k \text{ if } \Pr(C_k | X = x) = \max_{g \in C} \Pr(g | X = x)$$

Bayes Classifier

SUMMARY: WHEN EPE USES DIFFERENT LOSS

Loss Function	Estimator $\hat{f}(x)$
L_2 	EPE = $E_{X,Y}(Y-f(X))^2$ $\hat{f}(x) = E[Y\|X=x]$
L_1 	$\hat{f}(x) = \text{median}(Y\|X=x)$
0-1 	$\hat{f}(x) = \arg\max_Y P(Y\|X=x)$ (Bayes classifier / MAP)

Today : Naïve Bayes Classifier

- ✓ Why Bayes Classification – MAP Rule?
  - Review: Mean & Variance
  - Empirical Prediction Error, 0-1 Loss function for Bayes Classifier
- ✓ Naïve Bayes Classifier for Text document categorization
  - ✓ Bag of words representation
  - ✓ Multinomial vs. multivariate Bernoulli

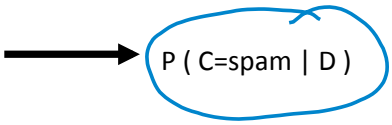
Text document classification, e.g. spam email filtering

- Input: document D
- Output: the predicted class C, c is from $\{c_1,...,c_L\}$

- Spam filtering Task: Classify email as *'Spam', 'Other'.*

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99



Text classification Tasks

- Input: document D
- Output: the predicted class C, c is from $\{c_1,...,c_L\}$

Text classification examples:

- Classify email as *'Spam', 'Other'.*
- Classify web pages as *'Student', 'Faculty', 'Other'*
- Classify news stories into topics *'Sports', 'Politics'* ⇒ Google News
- Classify business names by industry.
- Classify movie reviews as *'Favorable', 'Unfavorable', 'Neutral'* ⇒ Netflix
- *… and many more.*

Text Classification: Examples

- Classify shipment articles into one 93 categories
- An example category 'wheat'

ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS
BUENOS AIRES, Feb 26
Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, showing those for future shipments month, 1986/87 total and 1985/86 total to February 12, 1986, in brackets:
Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total 2,692.4 (4,161.0).
Maize Mar 48.0, total 48.0 (nil).
Sorghum nil (nil)
Oilseed export registrations were:
Sunflowerseed total 15.0 (7.9)
Soybean May 20.0, total 20.0 (nil)
The board also detailed export registrations for subproducts, as follows....

Representing text: a list of words

argentine, 1986, 1987, grain, oilseed, registrations, buenos, aires, feb, 26, argentine, grain, board, figures, show, crop, registrations, of, grains, oilseeds, and, their, products, to, february, ~~11~~, in, …

→ C

Common refinements: remove stopwords, stemming, collapsing multiple occurrences of words into one….

'Bag of words' representation of text

ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS
BUENOS AIRES, Feb 26
Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, showing those for future shipments month, 1986/87 total and 1985/86 total to February 12, 1986, in brackets:
Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total 2,692.4 (4,161.0).
Maize Mar 48.0, total 48.0 (nil).
Sorghum nil (nil)
Oilseed export registrations were:
Sunflowerseed total 15.0 (7.9)
Soybean May 20.0, total 20.0 (nil)
The board also detailed export registrations for sub-products, as follows....

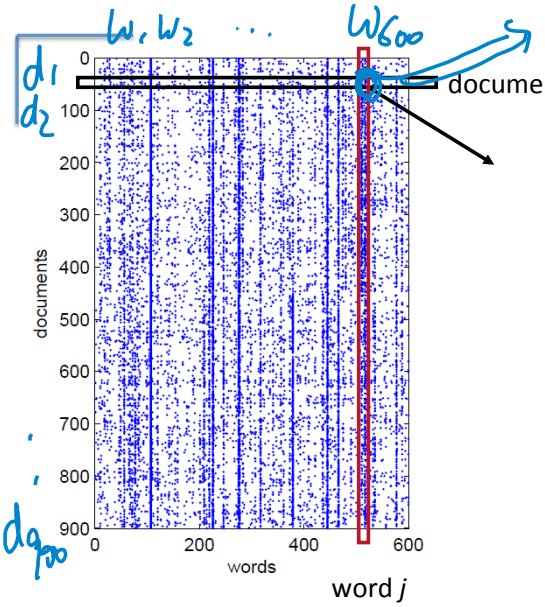
→

word	*frequency*
grain(s)	3
oilseed(s)	2
total	3
wheat	1
maize	1
soybean	1
tonnes	1
...	...

Bag of word representation:

Represent text as a vector of word *frequencies*.

Bag of words representation



X (i,j) = Frequency of word j in document i

many other choices, e.g. BM25 / TF-IDF / ..

A collection of documents

	X_1	X_2	X_3	C
S_1				
S_2				
S_3				
S_4				
S_5				
S_6				

Bag of words

- What simplifying assumption are we taking?

We assumed *word order* is not important.

Today : Naïve Bayes Classifier

- ✓ Why Bayes Classification – MAP Rule?
  - Review: Mean & Variance
  - Empirical Prediction Error, 0-1 Loss function for Bayes Classifier
- ✓ Naïve Bayes Classifier for Text document categorization
  - ✓ Bag of words representation
  - ✓ Multinomial vs. multivariate Bernoulli

'Bag of words' representation of text

$D = (w_1, w_2, \ldots, w_K)$

ARGENTINE 1986/87 GRAIN/OILSEED REGISTRATIONS
BUENOS AIRES, Feb 26
Argentine grain board figures show crop registrations of grains, oilseeds and their products to February 11, in thousands of tonnes, showing those for future shipments month, 1986/87 total and 1985/86 total to February 12, 1986, in brackets:
Bread wheat prev 1,655.8, Feb 872.0, March 164.6, total 2,692.4 (4,161.0).
Maize Mar 48.0, total 48.0 (nil).
Sorghum nil (nil)
Oilseed export registrations were:
Sunflowerseed total 15.0 (7.9)
Soybean May 20.0, total 20.0 (nil)
The board also detailed export registrations for sub-products, as follows....

word	*frequency*
grain(s)	3
oilseed(s)	2
total	3
wheat	1
maize	1
soybean	1
tonnes	1
...	...

$P(C|X) \propto P(C)\,P(X|C)$

$\Pr(D \mid C = c)$?

'Bag of words' representation of text

$\Pr(D \mid C = c)$? $D = (w_1, w_2, \ldots, w_K)$

Two Previous models

$\Pr(W_1 = true, W_2 = false \ldots, W_k = true \mid C = c)$

$\Pr(W_1 = n_1, W_2 = n_2, \ldots, W_k = n_k \mid C = c)$

34

Note: Two Models

- *Model 1*: Multivariate Bernoulli
  - One feature X_w for each word in dictionary
  - X_w = true in document d if w appears in d
  - Naive Bayes assumption:
    - Given the document's topic class label, appearance of one word in the document tells us nothing about chances that another word appears

$$\Pr(W_1 = true, W_2 = false..., W_k = true \mid C = c)$$

Model 1: Multivariate Bernoulli

$$P(w_1, w_2, .., w_K | C) = P(w_1|C)P(w_2|C) .. P(w_K|C)$$

word	*True/false*
grain(s)	True
oilseed(s)	True
total	True
wheat	True
chemical	False
...	...

- **Conditional Independence Assumption:** Features (word presence) are *independent* of each other given the class variable:
- Multivariate Bernoulli model is appropriate for binary feature variables

Model 2: Multinomial Naïve Bayes
- 'Bag of words' representation of text

word	*frequency*
grain(s)	3
oilseed(s)	2
total	3
wheat	1
maize	1
soybean	1
tonnes	1
...	...

$$\Pr(W_1 = n_1, \ldots, W_k = n_k \mid C = c)$$

Can be represented as a multinomial distribution.

Words = like colored balls, there are *K* possible type of them (i.e. from a dictionary of K words)

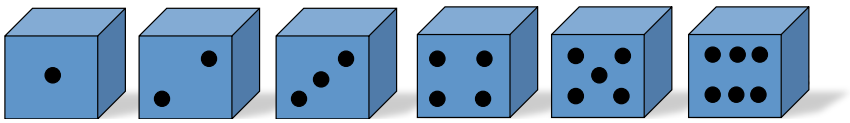
Document = contains N words, each word occurs n_i times (like a bag of N colored balls)

In a document class of 'wheat', "grain" is more likely. where as in a "hard drive" shipment class, the parameter for 'grain' is going to be smaller.

Multinomial distribution

- The multinomial distribution is a generalization of the binomial distribution.
- The binomial distribution counts successes of an event (for example, heads in coin tosses). K=2
- The parameters:
  - *N* (number of trials)
  - θ (the probability of success of the event)

- The multinomial counts the number of a set of events (for example, how many times each side of a die comes up in a set of rolls).
  - The parameters:
  - *N* (number of trials)
  - $\theta_1 .. \theta_k$ (the probability of success for each category) K=6



Multinomial Distribution

$P(D|c_i) = P(w_1, w_2 .. w_k | c)$

- $W_1, W_2, .. W_k$ are variables

Number of possible orderings of N balls

$$P(W_1 = n_1, ..., W_k = n_k \mid N, \theta_1, .., \theta_k) = \frac{N!}{n_1! n_2! .. n_k!} \theta_1^{n_1} \theta_2^{n_2} .. \theta_k^{n_k} \quad | c = c_i$$

order invariant selections

Note events are independent

$$\sum_{i=1}^{k} n_i = N \qquad \sum_{i=1}^{k} \theta_i = 1$$

A binomial distribution is the multinomial distribution with k=2 and θ_1

$\theta_2 = 1 - \theta_1$

Text Classification with Naïve Bayes Classifier

- Multinomial vs Multivariate Bernoulli?
- Multinomial model is almost always more effective in text applications!

Adapt From Manning' textCat tutorial

Experiment: Multinomial vs multivariate Bernoulli

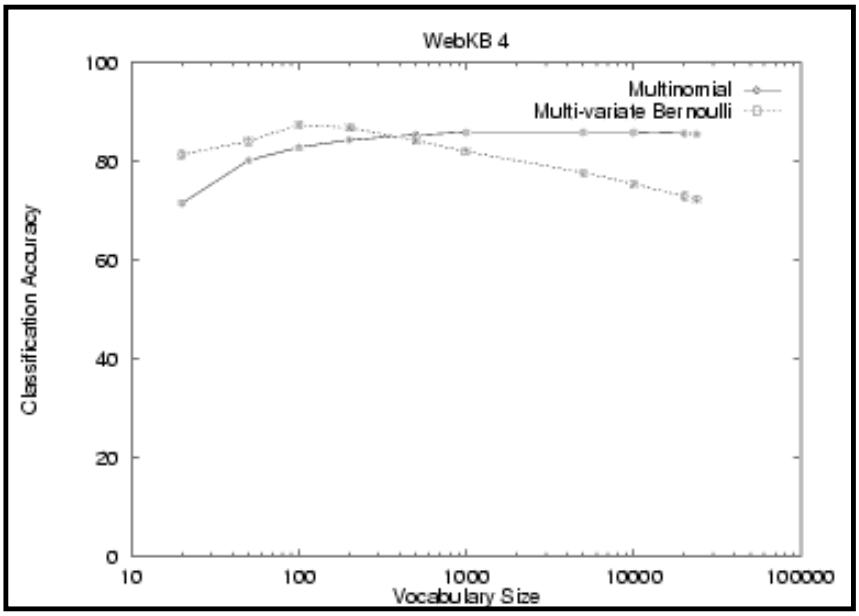
- M&N (1998) did some experiments to see which is better
- Determine if a university web page is {student, faculty, other_stuff}
- Train on ~5,000 hand-labeled web pages
  - Cornell, Washington, U.Texas, Wisconsin
- Crawl and classify a new site (CMU)

train

test

Adapt From Manning' textCat tutorial

Multinomial vs. multivariate Bernoulli



Today Recap: Naïve Bayes Classifier

- ✓ Why Bayes Classification – MAP Rule?
  - Review: Mean & Variance
  - Empirical Prediction Error, 0-1 Loss function for Bayes Classifier
- ✓ Naïve Bayes Classifier for Text document categorization
  - ✓ Bag of words representation
  - ✓ Multinomial vs. multivariate Bernoulli

References

- ❑ Prof. Andrew Moore's review tutorial
- ❑ Prof. Ke Chen NB slides
- ❑ Prof. Carlos Guestrin recitation slides
- ❑ Prof. Manning's textCat tutorial