

# UVA CS 4501 - 001 / 6501 – 007

## Introduction to Machine Learning and Data Mining

### Lecture 14: Naïve Bayes Classifier (cont.)

Yanjun Qi / Jane, , PhD

University of Virginia  
Department of  
Computer Science

10/17/14

1

## Where are we ? →

### Five major sections of this course

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

10/17/14

2

## 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

## Last Lecture Recap:

### Naïve Bayes Classifier

- ✓ Probability review
  - Structural properties, i.e., Independence, conditional independence
- ✓ Naïve Bayes Classifier
  - Spam email classification

## Review : Probability

- Joint  $\sum_x \sum_y P(X = x \cap Y = y) = 1$

- Conditional  $P(X = x | Y = y) = \frac{P(X = x \cap Y = y)}{P(Y = y)}$

- Marginal

$$\begin{aligned}
 P(X = x_i) &= \sum_j P(X = x_i \cap Y = y_j) \\
 &= \sum_j P(X = x_i | Y = y_j) P(Y = y_j)
 \end{aligned}$$

Marginal Probability (points to  $P(X = x_i)$ )  
 Joint Probability (points to  $P(X = x_i \cap Y = y_j)$ )  
 Conditional Probability (points to  $P(X = x_i | Y = y_j)$ )  
 Marginal Probability (points to  $P(Y = y_j)$ )  
*chain rule* (points to the multiplication in the second line)

10/17/14

## Review : Probability

- Independence

$$P(X = x \cap Y = y) = P(X = x)P(Y = y)$$

- Conditional independence

$$P(X = x \cap Y = y | Z = z) = P(X = x | Z = z)P(Y = y | Z = z)$$

10/17/14

6

# Review : Bayes' Rule

$$P(C, X) = P(C | X)P(X) = P(X | C)P(C)$$

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

Posterior

Prior

$P(C_1|x), P(C_2|x), \dots, P(C_L|x)$

$P(C_1), P(C_2), \dots, P(C_L)$

$X_1$	$X_2$	$X_3$	$C$

## A Dataset for classification

$$f : X \rightarrow C$$

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

$$P(C | X)$$

- **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  $\mathbf{x}$  with attributes  $(x_1, x_2, \dots, x_p)$ :
  - Goal is to predict class  $C$ .
  - Specifically, we want to find the value of  $C_i$  that maximizes  $p(C_i | x_1, x_2, \dots, x_p)$ .

- Bayes classification

$$P(C | \mathbf{X}) \propto P(\mathbf{X} | C)P(C) = P(X_1, \dots, X_p | C)P(C)$$

Difficulty: learning the joint probability  $P(X_1, \dots, X_p | C)$

10/17/14

9

## Naïve Bayes Classifier

Difficulty: learning the joint probability  $P(X_1, \dots, X_p | C)$

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

$$\begin{aligned} P(X_1, X_2, \dots, X_p | C) &= P(X_1 | X_2, \dots, X_p, C)P(X_2, \dots, X_p | C) \\ &= P(X_1 | C)P(X_2, \dots, X_p | C) \\ &= P(X_1 | C)P(X_2 | C) \cdots P(X_p | C) \end{aligned}$$

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

$$[P(x_1 | c^*) \cdots P(x_p | c^*)]P(c^*) > [P(x_1 | c) \cdots P(x_p | c)]P(c),$$

$$c \neq c^*, c = c_1, \dots, c_L$$

10/17/14

10

Adapt from Prof. Ke Chen NB slides



# Naïve Bayes Classifier

- Learning Phase

$$P(X_2|C_1), P(X_2|C_2)$$

Outlook	Play=Yes	Play=No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=Yes	Play=No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=Yes	Play=No
High	3/9	4/5
Normal	6/9	1/5

$$P(X_4|C_1), P(X_4|C_2)$$

Wind	Play=Yes	Play=No
Strong	3/9	3/5
Weak	6/9	2/5

3+3+2+2 [naïve assumption] \* 2 [two classes]= 20 parameters

$$P(\text{Play=Yes}) = 9/14 \quad P(\text{Play=No}) = 5/14$$

$$P(C_1), P(C_2), \dots, P(C_L)$$

10/17/14

13

# Naïve Bayes Assumption

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

If no naïve assumption

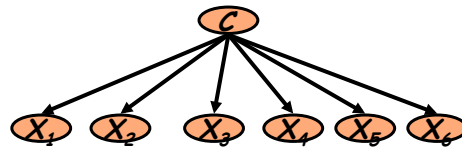
### Naïve Bayes Conditional Independence Assumption:

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

10/17/14

Adapt From Manning' textCat tutorial<sup>14</sup>

## Learning the Model



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

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

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

10/17/14

15

## Smoothing to Avoid Overfitting

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

# of values of feature  $X_i$

To make  
sum\_i (P(x\_i | C\_j)) = 1

10/17/14

Adapt From Manning' textCat tutorial<sup>16</sup>



## Smoothing to Avoid Overfitting

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

# of values of  $X_i$

- Somewhat more subtle version

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

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

extent of  
"smoothing"

10/17/14

17

## 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
- ✓ Multinomial naïve Bayes classifier

10/17/14

18

## Bayes Classifiers – MAP Rule

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

$$\begin{aligned}
 c_{MAP} &= \operatorname{argmax}_{c_j \in C} P(c_j | x_1, x_2, \dots, x_p) \quad \leftarrow \text{WHY?} \\
 &= \operatorname{argmax}_{c_j \in C} \frac{P(x_1, x_2, \dots, x_p | c_j) P(c_j)}{P(x_1, x_2, \dots, x_p)} \\
 &= \operatorname{argmax}_{c_j \in C} P(x_1, x_2, \dots, x_p | c_j) P(c_j)
 \end{aligned}$$

MAP = Maximum A posteriori Probability

10/17/14

Adapt From Carols' prob tutorial 19

## 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:  $E(X) = \int_{-\infty}^{+\infty} xf(x) dx$

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

10/17/14

Adapt From Carols' prob tutorial 20

## 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:

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

10/17/14

Adapt From Carols' prob tutorial

21

## Review: Mean and Variance of RV

- Mean

–  $E(X + Y) = E(X) + E(Y)$

–  $E(aX) = aE(X)$

– If X and Y are independent,  $E(XY) = E(X) \cdot E(Y)$

- Variance

–  $V(aX + b) = a^2 V(X)$

– If X and Y are independent,  $V(X + Y) = V(X) + V(Y)$

10/17/14

Adapt From Carols' prob tutorial

22

## Review: Mean and Variance of RV

- The conditional expectation of Y given X when the value of  $X = x$  is:

$$E(Y | X = x) = \int y * p(y | x) dy$$

- The Law of Total Expectation or Law of Iterated Expectation:

$$E(Y) = E[E(Y | X)] = \int E(Y | X = x) p_X(x) dx$$

## Review: Continuous Random Variables

- Probability density function (pdf) instead of probability mass function (pmf)
- 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

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

10/17/14

25

## 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
- ✓ Multinomial naïve Bayes classifier

10/17/14

26

## 0-1 LOSS for Classification

- Procedure for categorical output variable  $C$
- Frequently, 0-1 loss function used:  $L(k, \ell) = 1 - \delta_{kl}$
- $L(k, \ell)$  is the price paid for misclassifying an element from class  $C_k$  as belonging to class  $C_\ell$

10/17/14

## Expected prediction error (EPE)

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

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

Consider sample population distribution

- Pointwise minimization suffices

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

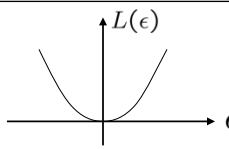
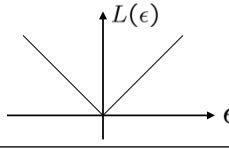
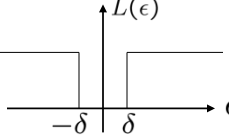
- → simply

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

Bayes Classifier

10/17/14

## SUMMARY: WHEN EPE USES DIFFERENT LOSS


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

10/17/14

Yanjun Qi / UVA CS 4501-01-6501-07

Yanjun Qi / UVA CS 4501-01-6501-07

## 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
  - ✓ Multinomial naïve Bayes classifier

10/17/14

30

## Text document classification, e.g. spam email filtering

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



$P(C = \text{spam} \mid D)$

## Text classification

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

### Text classification examples:

- Classify **email** as *'Spam'*, *'Other'*.
- Classify **web pages** as *'Student'*, *'Faculty'*, *'Other'*
- Classify **news stories** into topics *'Sports'*, *'Politics'*..
- Classify **business names** by industry.
- Classify **movie reviews** as *'Favorable'*, *'Unfavorable'*, *'Neutral'*
- ... 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,  
 registration, buenos, aires, feb, 26, argentine,  
 grain, board, figures, show, crop, registration,  
 of, grains, oilseeds, and, their, products, to,  
 february, 11, in, ...

→ C

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

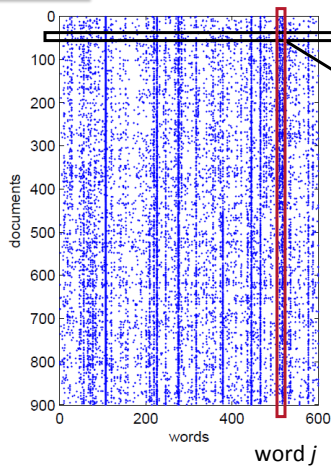
# 'Bag of words' representation of text

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



$Frequency(i,j) = j \text{ in document } i$

A collection of documents

	$X_1$	$X_2$	$X_3$	C
$S_1$				
$S_2$				
$S_3$				
$S_4$				
$S_5$				
$S_6$			36	

# Bag of words

- What simplifying assumption are we taking?

We assumed *word order* is not important.



## ‘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
...	...

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

Two models

$\Pr(W_1 = n_1, \dots, W_k = n_k | C = c)$

$\Pr(W_1 = true, W_1 = false, \dots, W_k = true | C = c)$

## 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
  - ✓ Multinomial naïve Bayes classifier

10/17/14

39

## Note: Two Models

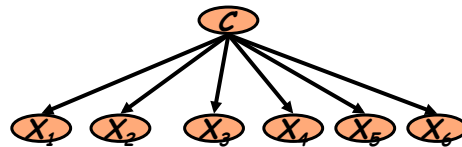
- **Model 1: Multivariate Bernoulli**
  - One feature  $X_w$  for each word in dictionary
  - $X_w = \text{true}$  in document  $d$  if  $w$  appears in  $d$
  
  - Naive Bayes assumption:
    - Given the document's topic class label, appearance of one word in the document tells us nothing about chances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

$$\Pr(W_1 = \text{true}, W_1 = \text{false} \dots, W_k = \text{true} \mid C = c)$$

10/17/14

Adapt From Manning' textCat tutorial<sup>40</sup>

## Model 1: Multivariate Bernoulli



- **Conditional Independence Assumption:** Features (word presence) are *independent* of each other given the class variable:

$$P(X_1, \dots, X_5 | C) = P(X_1 | C) \cdot P(X_2 | C) \cdot \dots \cdot P(X_5 | C)$$

- Multivariate Bernoulli model is appropriate for **binary feature variables**

10/17/14

Adapt From Manning' textCat tutorial <sup>41</sup>

## 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
...	...

$$\Pr(W_1 = n_1, \dots, W_k = n_k | C = c)$$

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)

The multinomial distribution of words is going to be different for different document class.

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.

10/17/14

42

## Text Classification with Naïve Bayes Classifier

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

10/17/14

Adapt From Manning' textCat tutorial<sup>43</sup>

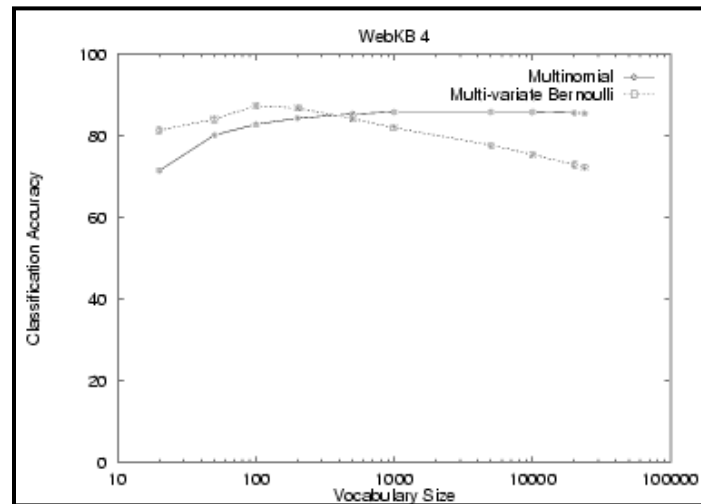
### Experiment: Multinomial vs multivariate Bernoulli

- 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)

10/17/14

Adapt From Manning' textCat tutorial<sup>44</sup>

## Multinomial vs. multivariate Bernoulli



10/17/14

Adapt From Manning' textCat tutorial <sup>45</sup>

## Today : Naïve Bayes Classifier

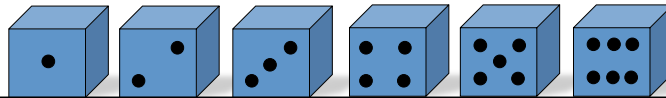
- ✓ Why Bayes Classification – MAP Rule?
  - Review: Mean & Variance
  - Empirical Prediction Error, 0-1 Loss function for Bayes Classifier
  
- ✓ Naïve Bayes Classifier for Text document categorization
  - ✓ Bag of words representation
  - ✓ Multinomial vs. multivariate Bernoulli
  - ✓ Multinomial naïve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

10/17/14

46

# 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).
- The parameters:
  - $N$  (number of trials)
  - $\theta$  (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)
    - $\theta_1 \dots \theta_k$  (the probability of success for each category)



10/17/14

# Multinomial Distribution

- $W_1, W_2, \dots, W_k$  are variables

$$P(W_1 = n_1, \dots, W_k = n_k \mid N, \theta_1, \dots, \theta_k) = \frac{N!}{n_1! n_2! \dots n_k!} \theta_1^{n_1} \theta_2^{n_2} \dots \theta_k^{n_k}$$

Number of possible orderings of N balls  $\nearrow$   
 $\nwarrow$  order invariant selections  
 $\searrow$  Note events are independent

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

A binomial distribution is the multinomial distribution with  $k=2$  and  $\theta_1, \theta_2 = 1 - \theta_2$

10/17/14

48



## Multinomial Distribution – a data generation process

- ✓ From a **box** you pick colored balls with  $k$  possible colors, replacing the extracted ball after each draw.
- ✓ Totally you did  $N$  independent draws randomly, i.e. totally you have  $N$  balls and put into your **bag**.
- ✓ Let probability of picking a ball of color  $i$  is  $\theta_i$
- ✓ For each color  $\theta_1, \dots, \theta_k$ , 
$$\sum_{i=1}^k \theta_i = 1$$
- ✓  $W_i$  be the random variable denoting the number of balls selected in color  $i$ , which can take values in  $\{1 \dots N\}$ . 
$$\sum_{i=1}^k W_i = N$$

YanJun Qi / UVA 6504501-01-6501-07

## Estimate parameter for Multinomial

- Now suppose  $X$  can have the values  $1, 2, \dots, K$   
(For example a die has  $K=6$  sides)
- We want to learn the parameters  $\theta_1, \theta_2, \dots, \theta_K$

### Sufficient statistics:

- ◆  $N_1, N_2, \dots, N_K$  - the number of times each outcome is observed

**Likelihood function:** 
$$L_D(\theta) = \prod_{k=1}^K \theta_k^{N_k}$$

**MLE:** 
$$\hat{\theta}_k = \frac{N_k}{\sum_{\ell} N_{\ell}}$$

10/17/14

## Parameter estimation

- **Multivariate Bernoulli model:**

$$\hat{P}(X_w = t | c_j) = \begin{array}{l} \text{fraction of documents of topic } c_j \\ \text{in which word } w \text{ appears} \end{array}$$

- **Multinomial model:**

$$\hat{P}(X_i = w | c_j) = \begin{array}{l} \text{fraction of times in which} \\ \text{word } w \text{ appears} \\ \text{across all documents of topic } c_j \end{array}$$

- Can create a mega-document for topic  $j$  by concatenating all documents on this topic
- Use frequency of  $w$  in mega-document

10/17/14

Adapt From Manning' textCat tutorial<sup>51</sup>

## Naïve Bayes: Learning Algorithm

- From training corpus, extract *Vocabulary*
- Calculate required  $P(c_j)$  and  $P(x_k | c_j)$  terms
  - For each  $c_j$  in  $C$  do
    - $docs_j \leftarrow$  subset of documents for which the target class is  $c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- $Text_j \leftarrow$  single document containing all  $docs_j$

- for each word  $x_k$  in *Vocabulary*

- $n_k \leftarrow$  number of occurrences of  $x_k$  in  $Text_j$

$$P(x_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

10/17/14

Adapt From Manning' textCat tutorial<sup>52</sup>

## Naïve Bayes: Classifying

- positions ← all word positions in current document which contain tokens found in *Vocabulary*
- Return  $c_{NB}$ , where

$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

Easy to implement,  
no need to construct  
bag-of-words vector  
explicitly !!!

10/17/14

Adapt From Manning' textCat tutorial <sup>53</sup>

## Naive Bayes: Time Complexity

- **Training Time:**  $O(|D|L_d + |C||V|)$ 
  - where  $L_d$  is the average length of a document in  $D$ .
  - Assumes  $V$  and all  $D_i$ ,  $n_i$ , and  $n_{ij}$  pre-computed in  $O(|D|L_d)$  time during one pass through all of the data.
  - Generally just  $O(|D|L_d)$  since usually  $|C||V| < |D|L_d$  ← Why?
- **Test Time:**  $O(|C| L_t)$ 
  - where  $L_t$  is the average length of a test document.
  - **Very efficient overall**, linearly proportional to the time needed to just read in all the data.
  - Plus, **robust** in practice

Adapt From Manning' textCat tutorial <sup>54</sup>

## Underflow Prevention: log space

- Multiplying lots of probabilities, which are between 0 and 1, can result in floating-point underflow.
- Since  $\log(xy) = \log(x) + \log(y)$ , it is better to perform all computations *by summing logs of probabilities rather than multiplying probabilities*.
- Class with highest final un-normalized log probability score is still the most probable.

$$c_{NB} = \operatorname{argmax}_{c_j \in C} \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i | c_j)$$

- Note that model is now just max of sum of weights...

## Naive Bayes is Not So Naive

- **Naïve Bayes: First and Second place in KDD-CUP 97 competition, among 16 (then) state of the art algorithms**

Goal: Financial services industry direct mail response prediction model: Predict if the recipient of mail will actually respond to the advertisement – 750,000 records.

- **Robust to Irrelevant Features**  
Irrelevant Features cancel each other without affecting results  
Instead Decision Trees can **heavily** suffer from this.
- **Very good in domains with many equally important features**  
Decision Trees suffer from *fragmentation* in such cases – especially if little data
- **A good dependable baseline for text classification (but not the best)!**
- **Optimal if the Independence Assumptions hold:** If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- **Very Fast:** Learning with one pass of counting over the data; testing linear in the number of attributes, and document collection size
- **Low Storage requirements**

## References

- Prof. Chris Manning' textCat review tutorial
- Prof. Carlos Guestrin recitation slides
- Prof. Ke Chen NB slides