## CS 6120/CS4120: Natural Language Processing

Instructor: Prof. Lu Wang College of Computer and Information Science Northeastern University Webpage: www.ccs.neu.edu/home/luwang

1

3

#### Outline

- Text Categorization/Classification
- Naïve Bayes
- Evaluation

2

## Positive or negative movie review?

- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot
- twists • this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.

#### Male or female author?

- By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
- Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

4

## Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- •...

5

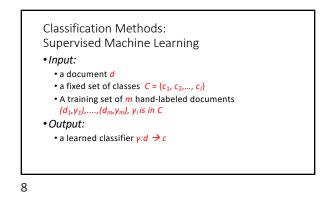
Text Classification: definition

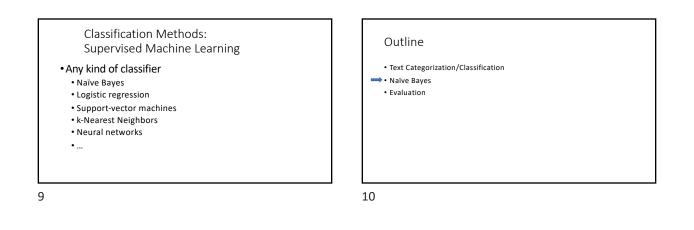
- •Input:
  - a document *d*
  - a fixed set of classes  $C = \{c_1, c_2, ..., c_J\}$
- *Output*: a predicted class  $c \in C$

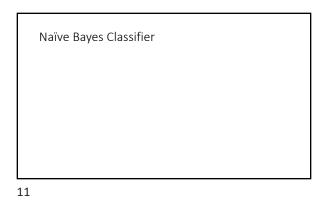
#### Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
   spam: black-list-address OR ("dollars" AND "have been selected")
- Accuracy can be high
- If rules carefully refined by expert
- But building and maintaining these rules is expensive

7

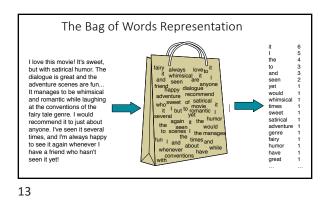


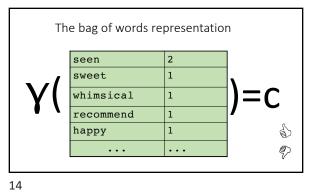


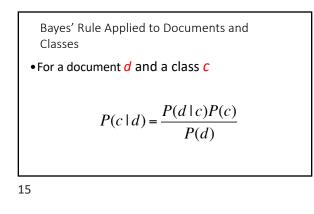


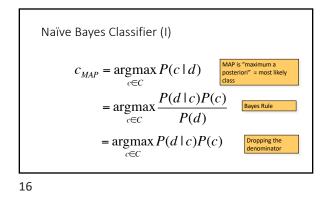
Naïve Bayes Intuition

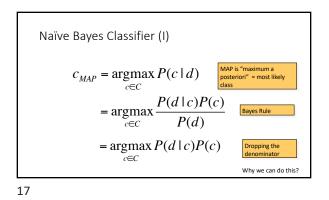
- Simple ("naïve") classification method based on Bayes rule
- Relies on very simple representation of document • Bag of words

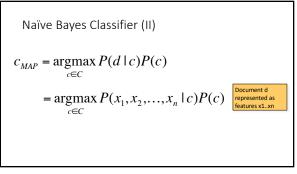


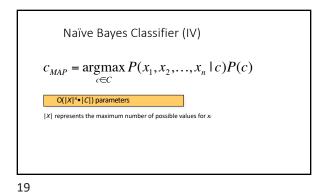


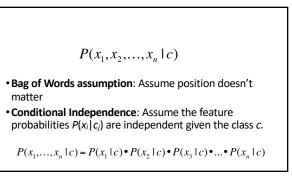


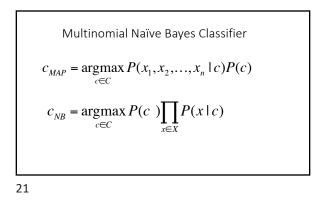




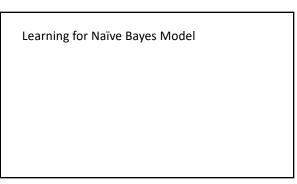


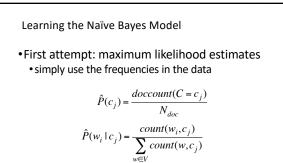


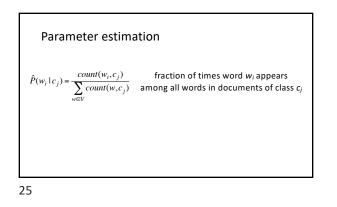


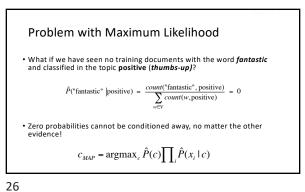


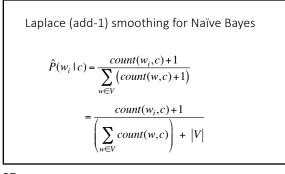
Applying Multinomial Naive Bayes Classifiers to Text Classification positions  $\leftarrow$  all word positions in test document  $c_{NB} = \underset{c_{j} \in C}{\operatorname{argmax}} P(c_{j}) \prod_{i \in positions} P(x_{i} | c_{j})$ 22



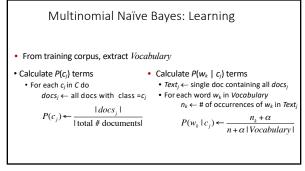




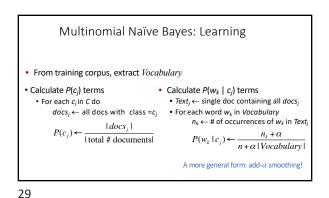


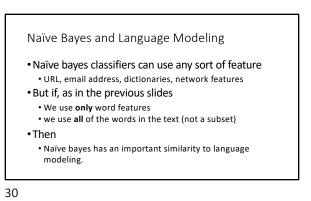




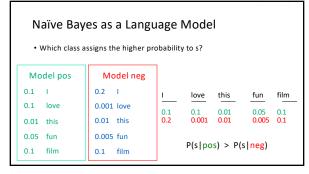


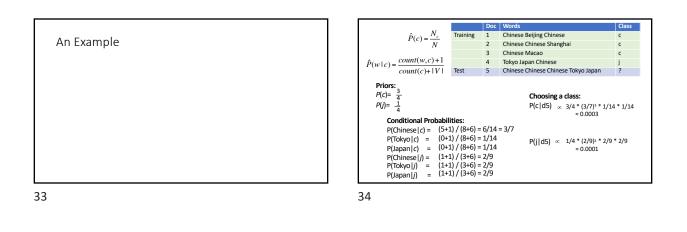


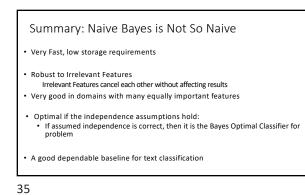




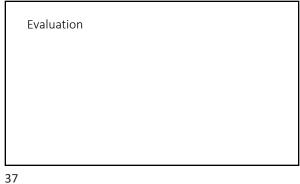
Ead	Each class = a unigram language model								
• As	<ul> <li>Assigning each word: P(word   c)</li> </ul>								
• •	<ul> <li>Assigning each sentence: P(sentence c)=Π P(word c)</li> </ul>								
1 1	ssigning each senter	ice. r (sei	litence ju	)=11 F (VV	siu(c)				
Class	pos								
0.1	i					<i>c</i>			
0.1	love	<u> </u>	love	this	fun	film			
0.1	1046	0.1	0.1	0.01	0.05	0.1			
0.01	this	0.1	0.1	0.01	0.05	0.1			
0.05	fun								
		P(sentence   pos) = 0.0000005							
0.1	film								









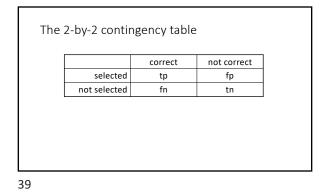


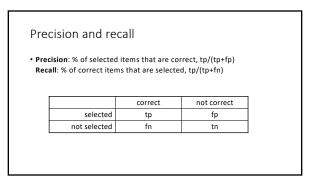
The 2-by-2 contingency table (or confusion matrix)

	correct	not correct		
selected	tp (true positive)	fp (false positive)		
not selected	fn (false negative)	tn (true negative)		

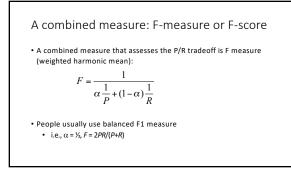
For example, • Which set of documents are related to the topic of NLP? • Which set of documents are written by Shakespeare?

38





40



41

Text Classification Evaluation

## More Than Two Classes: Sets of binary classifiers

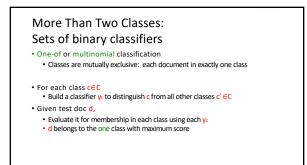
# • Dealing with any-of or multivalue classification

A document can belong to 0, 1, or >1 classes.

#### For each class c∈C

- Build a classifier  $\gamma_c$  to distinguish c from all other classes  $c'\in C$  Given test doc d,
  - Evaluate it for membership in each class using each  $\gamma_c$ • d belongs to any class for which  $\gamma_c$  returns true

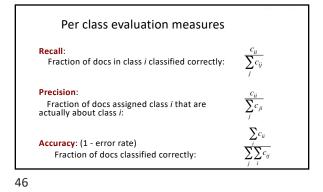
43

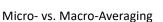


44

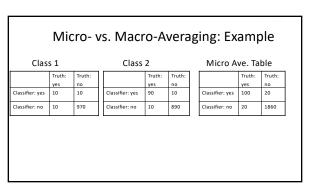
Confusion matrix c For each pair of classes $$ how many documents from $c_1$ were								
<ul> <li>incorrectly assigned to c<sub>2</sub>?</li> <li>c<sub>3,2</sub>: 90 wheat documents incorrectly assigned to poultry</li> </ul>								
Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade		
True UK	95	1	13	0	1	0		
True poultry	0	1	0	0	0	0		
True wheat	10	90	0	1	0	0		
True coffee	0	0	0	34	3	7		
True interest	0	1	2	13	26	5		
True trade	0	0	2	14	5	10		

45





- If we have more than one class, how do we combine multiple performance measures into one quantity?
- Macroaveraging: Compute performance for each class, then average.
- Microaveraging: Collect decisions for all classes, compute contingency table, evaluate.



Class 1			Class 2			Micro Ave. Table		
	Truth:	Truth:		Truth:	Truth:		Truth:	Truth:
	yes	no		yes	no		yes	no
Classifier: yes	10	10	Classifier: yes	90	10	Classifier: yes	100	20
Classifier: no	10	970	Classifier: no	10	890	Classifier: no	20	1860
Macroaveraged precision: (0.5 + 0.9)/2 = 0.7     Microaveraged precision: 100/120 = .83								

Development Test Sets and Cross-validation								
	Training set	ing set Development/tuning/held-out Set Test So						
N	Metric: P/R/F1 or Accuracy							
Cross-validation over multiple splits • Handle sampling errors from different datasets					Cev Tes	t		
Pool results over each split     Compute pooled dev set performance				Training	Set	Dev Test		
					Test Training S			
Test Set								