

# Bayes and Naïve Bayes

## Text Classification

J.Nerbonne, using slides by Dan Jurafsky

# Conditional probabilities

- Conditional probability of  $b$  given  $a$  is the probability of  $b$  in all those cases in which  $a$  holds.  $P(b|a)$
- Probability of voting for candidates on the left is higher among women
  - $P(\text{left} | \text{woman}) > P(\text{left} | \text{man})$
- Probability of common noun higher after an article
  - $P(\text{CN}_{i+1} | \text{Art}_i) > P(\text{Cn}_{i-1} | \sim \text{Art}_i)$
- Defined:  $\mathbf{p(B|A) = P(A,B)/P(A)}$



Thomas Bayes (1701-1761)

# Text Classification and Naïve Bayes

## Naïve Bayes (I)

# A problem in diagnosis

- You feel tired, weak and your joints are sore. You've also noticed some bleeding in your gums.
- A diagnostic handbook names these as symptoms of SCURVY, a dangerous disease normally caused by vitamin C deficiency.
- But when you tell this to your doctor, she examines you and then recommends getting more rest and staying warm. You're probably getting a flu.
- Is this irresponsible?

# What are the factors?

- Flu is very common, scurvy isn't, 90% vs. 10%
- In Bayesian terms, these are the PRIOR PROBABILITIES, and the Bayesian approach emphasizes their importance
- The symptoms are not uncommon in flu (20%) and very common in scurvy (80%)

# Thinking visually



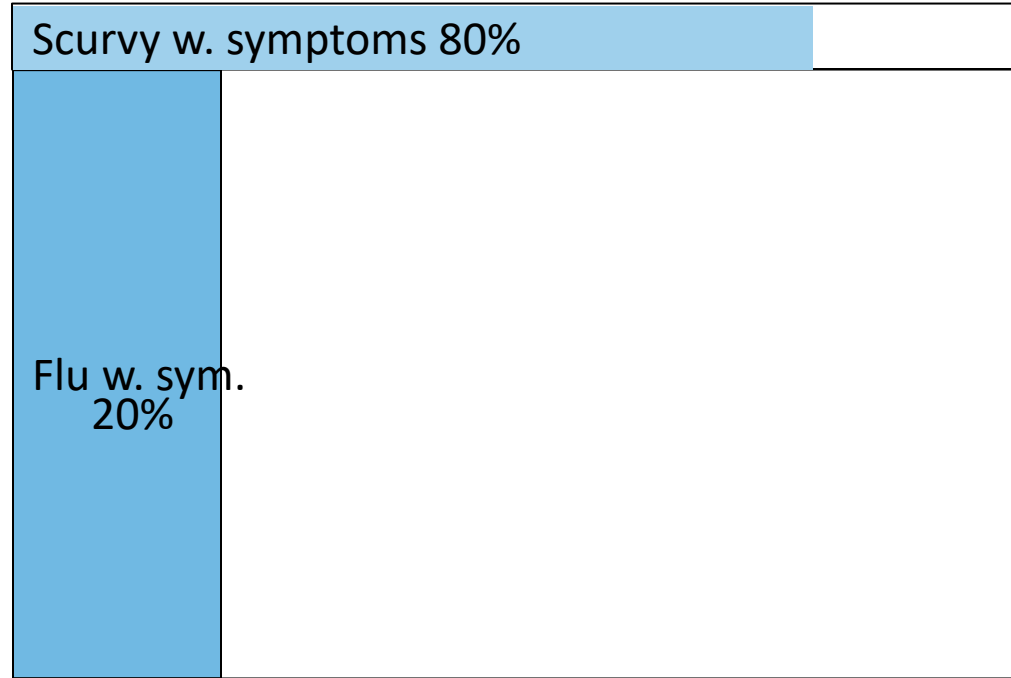
Scurvy sufferers

Flu sufferers

# Thinking visually

$$0.8 \times 0.1 = 0.08$$

$$0.2 \times 0.9 = 0.18$$



# Iteration is powerful

- We update our priors to 0.08 for scurvy and 0.18 for flu
- Then you insist on a test for vitamin C deficiency. If the test is positive, then there's a 75% of scurvy and a 25% of flu
- Posterior probabilities are then 0.06 for scurvy and 0.045 for flu!
- Most ML uses of Bayes involve retraining after adjusting priors



# Simple derivation of Bayes's Rule

- Recall definition of conditional probability

$$P(a|c) = P(a, c)/P(c)$$

- Note that therefore  $P(a|c)P(c) = P(a, c)$
- Note that  $P(c|a) = P(c, a)/P(a)$ , thus  $P(c|a)P(a) = P(c, a)$
- Since  $P(a, c) = P(c, a)$ (joint probability):

$$P(a|c)P(c) = P(c|a)P(a)$$

$$P(c|a) = P(a|c)P(c)/P(a)$$

- Bayes's rule:  $P(c|a) = P(a|c)P(c)/P(a)$

# Document Classification

- Apply to document classification:

$$P(c|d) = P(d|c)P(c)/P(d)$$

- Which class  $c$  is most likely?
- Just compare the  $P(c_i|d)$  for all  $c_i$ ; the largest probability is the most likely
- Called the MAXIMUM A POSTERIORI (MAP) hypothesis

$$C_{MAP} = \operatorname{argmax}_{c_i \in C} P(c_i | d)$$

# Naïve Bayes Classifier (I)

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

$d$  is constant in comparison, so we can drop the denominator,

# Bayesian terminology

- $P(c|a) = P(a|c)P(c)/P(a)$  – Bayes's rule
- Posterior-Prob (hypothesis|data)  
= LIKELIHOOD (data|hypo) X PRIOR-PROB (hypo) / Prob.(evidence)
- In medical diagnosis, the diagnosis is the hypothesis
  - Patient w. headaches,  $\nexists$ tumor (LIKELIHOOD headaches high, given tumor)
  - But tumors much less frequent than dehydration, trauma, fatigue, ...
    - Prior probability of tumor low
  - Therefore diagnosis – no tumor (initially)

# Bayes outside CL

- Attention to prior probabilities is attractive!
- Popular in medical testing
- Used by the insurance industry (differential rates for young/old)
- Used increasingly as alternative to frequentist hypothesis testing
  - Dissatisfaction with “null hypothesis significance testing”
  - “Small p-values don’t mean small chance of being wrong”
- Used widely in machine learning (more on this)

# Naïve Bayes for Document Classifiers (I)

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

$d$  is constant in comparison, so we can drop the denominator,

# Naïve Bayes Intuition

- Simple (“naïve”) classification method based on Bayes rule
- Relies on very simple representation of document
  - BAG OF WORDS

# The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...



# Which words are in the bag?


- Information retrieval uses only content words, no very frequent (function) words
- Authorship attribution mostly uses only very frequent words
- Spam detection uses everything!
- Language detection uses n-grams of characters (letters)
- Gender detection (profiling), sentiment analysis, ...

# The bag of words representation

$Y($

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

$) = C$


# Combinations of words!

- We're now representing documents as combinations of features, e.g., the fact that the word *phoneme* occurs four times, *phonotactic* twice, and *hate* zero times
- This makes gathering statistics difficult, since there may be **no** document where *phoneme* occurs 4 times, *phonotactic* twice, and *hate* zero
- ... no real solution to this problem, but it turns out not to be devastating

# Naïve Bayes for Document Classifiers (II)

$$\begin{aligned}C_{MAP} &= \operatorname{argmax}_{c_i \in C} P(c_i | d) \\&= \operatorname{argmax}_{c_i \in C} P(d | c_i) P(c_i) \\&= \operatorname{argmax}_{c_i \in C} P(x_1, x_2, \dots, x_n | c_i) P(c_i)\end{aligned}$$

Document d  
represented as  
features(words)  
x1..xn

This just recaps the last slide more mathematically.

# Naïve Bayes for Document Classifiers (III)

$$C_{MAP} = \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c) P(c)$$

$O(|X|^n \cdot |C|)$  parameters

How often does this class occur?

Could only be estimated w. a very large amount of data

We can just count the relative frequencies in a corpus

# Solution: Naïve Bayes Independence Assumptions

$$P(x_1, x_2, \dots, x_n | c) P(c)$$

- **Bag-of-Words assumption:** Assume position doesn't matter
- **Conditional Independence:** Assume the feature probabilities  $P(x_i | c_j)$  are independent given the class  $c$ .

$$P(x_1, x_2, \dots, x_n | c) = P(x_1 | c) \times P(x_2 | c) \times \dots \times P(x_n | c)$$

- Especially conditional independence is wrong. Why?
- But naïve Bayes often works quite well!

# Text Classification and Naïve Bayes

Naïve Bayes:  
Learning

# Learning the Naïve Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{doc}}$$

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$



# Problem with Maximum Likelihood

- What if we have seen no training documents with the word *fantastic* and classified in the topic **positive** (*thumbs-up*)?

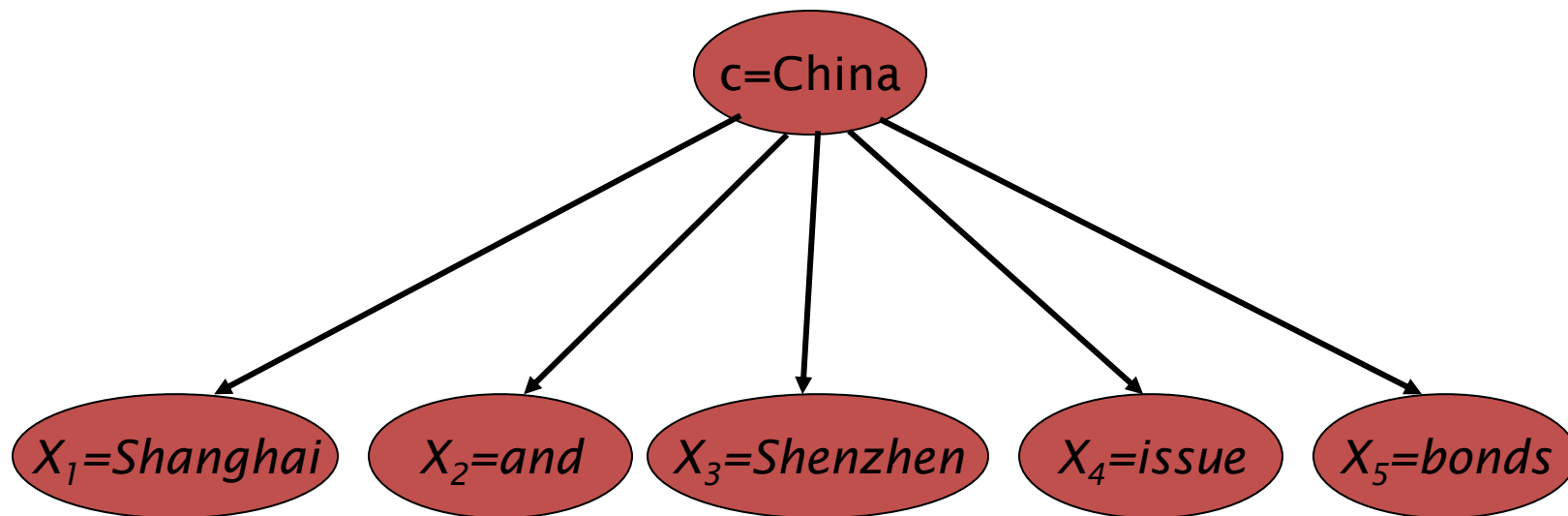
$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{\text{count}(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \operatorname{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

- Then we need to consider smoothing estimations, just as in language models -- +1, Laplace, ...

# Generative Model for Multinomial Naïve Bayes



# Naïve Bayes and Language Modeling

- Naïve Bayes classifiers can use any sort of feature
  - URL, email address, dictionaries, network features
- But if, as in the previous slides
  - We use **only** word features
  - We use **all** of the words in the text (not a subset)
- Then
  - Naïve Bayes can be seen as similar to language modeling.
    - (LSP, 4.6)

# Each class = a unigram language model

- Assigning each word:  $P(\text{word} \mid c)$
- Assigning each sentence:  $P(s \mid c) = \prod P(\text{word} \mid c)$

Class *pos*

0.1	I	<u>I</u>	<u>love</u>	<u>this</u>	<u>fun</u>	<u>film</u>
0.1	love	0.1	0.1	.05	0.01	0.1
0.01	this					
0.05	fun					
0.1	film					

$$P(s \mid \text{pos}) = 0.00000005$$

# Naïve Bayes as a Language Model

- Which class assigns the higher probability to s?

Model pos	
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model neg	
0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

<u>I</u>	<u>love</u>	<u>this</u>	<u>fun</u>	<u>film</u>
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$$P(s|\text{pos}) > P(s|\text{neg})$$

# Text Classification and Naïve Bayes

Multinomial Naïve  
Bayes: A Worked  
Example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w | c) = \frac{\text{count}(w, c) + 1}{\text{count}(c) + |V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

**Priors:**

$$P(c) = \frac{3}{4}$$

$$P(j) = \frac{1}{4}$$

**Choosing a class:**

$$P(c | d_5) \propto \frac{3}{4} * \left(\frac{3}{7}\right)^3 * \frac{1}{14} * \frac{1}{14} \approx 0.0003$$

**Conditional Probabilities:**

$$P(\text{Chinese} | c) = \frac{(5+1)}{(8+6)} = \frac{6}{14} = \frac{3}{7}$$

$$P(\text{Tokyo} | c) = \frac{(0+1)}{(8+6)} = \frac{1}{14}$$

$$P(\text{Japan} | c) = \frac{(0+1)}{(8+6)} = \frac{1}{14}$$

$$P(\text{Chinese} | j) = \frac{(1+1)}{(3+6)} = \frac{2}{9}$$

$$P(\text{Tokyo} | j) = \frac{(1+1)}{(3+6)} = \frac{2}{9}$$

$$P(\text{Japan} | j) = \frac{(1+1)}{(3+6)} = \frac{2}{9}$$

$$P(j | d_5) \propto \frac{1}{4} * \left(\frac{2}{9}\right)^3 * \frac{2}{9} * \frac{2}{9} \approx 0.0001$$

# Naïve Bayes in Spam Filtering

- SpamAssassin Features:
  - Mentions Generic Viagra
  - Online Pharmacy
  - Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
  - Phrase: impress ... girl
  - From: starts with many numbers
  - Subject is all capitals
  - HTML has a low ratio of text to image area
  - One hundred percent guaranteed
  - Claims you can be removed from the list
  - 'Prestigious Non-Accredited Universities'
  - [http://spamassassin.apache.org/tests\\_3\\_3\\_x.html](http://spamassassin.apache.org/tests_3_3_x.html)



# Summary: Naive Bayes is Not So Naive

- Very Fast, low storage requirements
- Robust to Irrelevant Features
  - Irrelevant Features cancel each other without affecting results
- Very good in domains with many equally important features
  - Decision Trees suffer from *fragmentation* in such cases – especially if little data
- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
  - **But there are other classifiers that give better accuracy**