Natural Language Processing

Naïve Bayes and Word Sense Disambiguation

Alla Rozovskaya

Lexical ambiguity

- Black **sea bass**, which are often called "rock **bass**," prefer natural, hard bottom areas and limestone ledges.
- The four major instruments in the string family, the violin, the viola, the cello and the double **bass**, are built the same way.

Word Sense Disambiguation

- The task of selecting the correct sense for a word
- Is potentially useful in many applications
  - E.g. Machine translation
  - Information retrieval

Supervised learning for WSD

- **Supervised** machine learning algorithms
- We need hand-labeled data
  - A set of example sentences where each ambiguous word is labeled with the corresponding word sense

What is Learning

- The Badges Game……
  - This is an example of the key learning protocol: supervised learning

Slide adopted from Dan Roth

Training data

+ Naoki Abe
- Myriam Abramson
+ David W. Aha
+ Kamal M. Ali
- Eric Allender
+ Dana Angluin
- Chidanand Apte
+ Minoru Asada
+ Lars Asker
+ Javed Aslam
+ Jose L. Balcazar
- Cristina Baroglio

+ Peter Bartlett
- Eric Baum
+ Welton Becket
- Shai Ben-David
+ George Berg
+ Neil Berkman
+ Malini Bhandaru
+ Bir Bhanu
+ Reinhard Blasig
- Avrim Blum
- Anselm Blumer
+ Justin Boyan

+ Carla E. Brodley
+ Nader Bshouty
- Wray Buntine
- Andrey Burago
+ Tom Bylander
+ Bill Byrne
- Claire Cardie
+ John Case
+ Jason Catlett
- Philip Chan
- Zhixiang Chen
- Chris Darken

Slide adopted from Dan Roth

The Badges game

- + Naoki Abe
- - Eric Baum

- Conference attendees to the 1994 Machine Learning conference were given name badges labeled with + or −.

- What function was used to assign these labels?

Slide adopted from Dan Roth

Raw test data

Gerald F. DeJong
Chris Drummond
Yolanda Gil
Attilio Giordana
Jiarong Hong
J. R. Quinlan
Priscilla Rasmussen
Dan Roth
Yoram Singer
Lyle H. Ungar

Slide adopted from Dan Roth

Labeled test data

+ Gerald F. DeJong
- Chris Drummond
+ Yolanda Gil
- Attilio Giordana
+ Jiarong Hong
- J. R. Quinlan
- Priscilla Rasmussen
+ Dan Roth
+ Yoram Singer
- Lyle H. Ungar

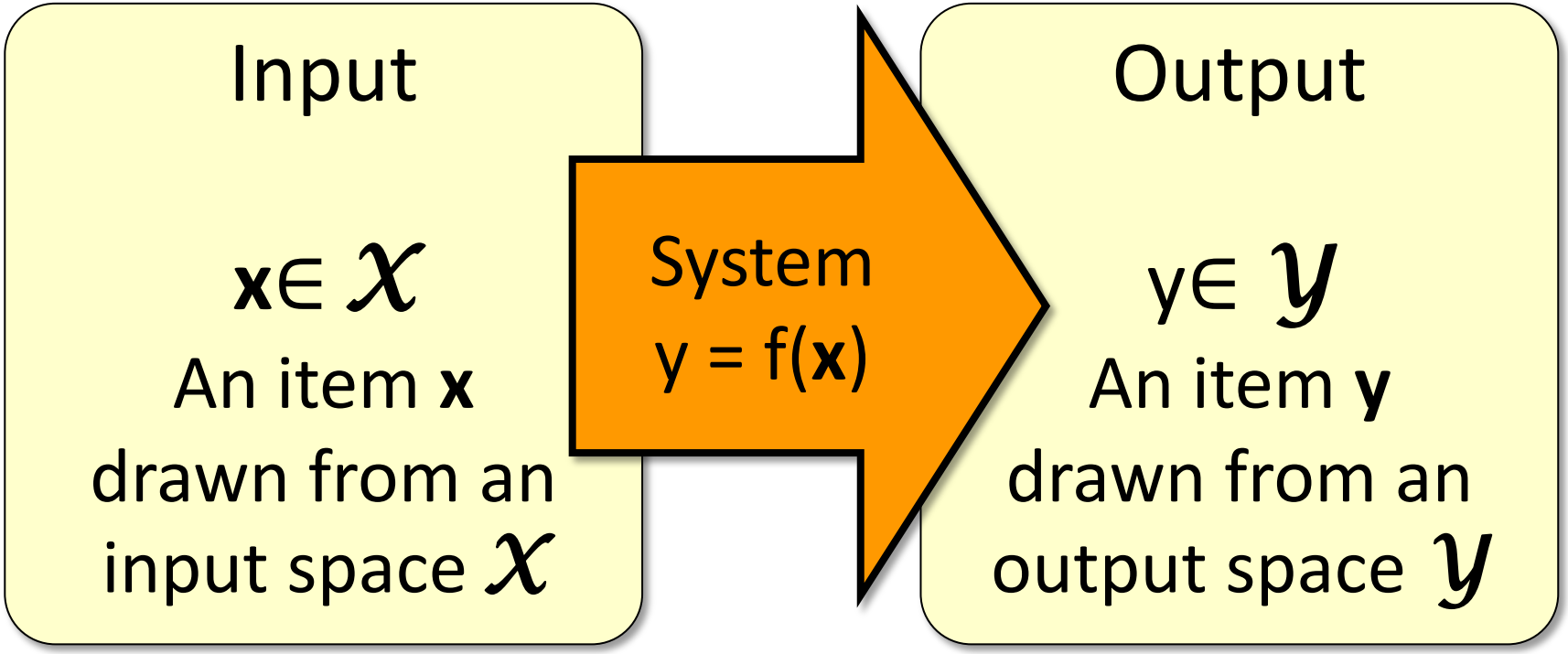
Slide adopted from Dan Roth

What is Learning

- The Badges Game……
  - This is an example of the key learning protocol: supervised learning
- First question: Are you sure you got it right?
  - Why?

Slide adopted from Dan Roth

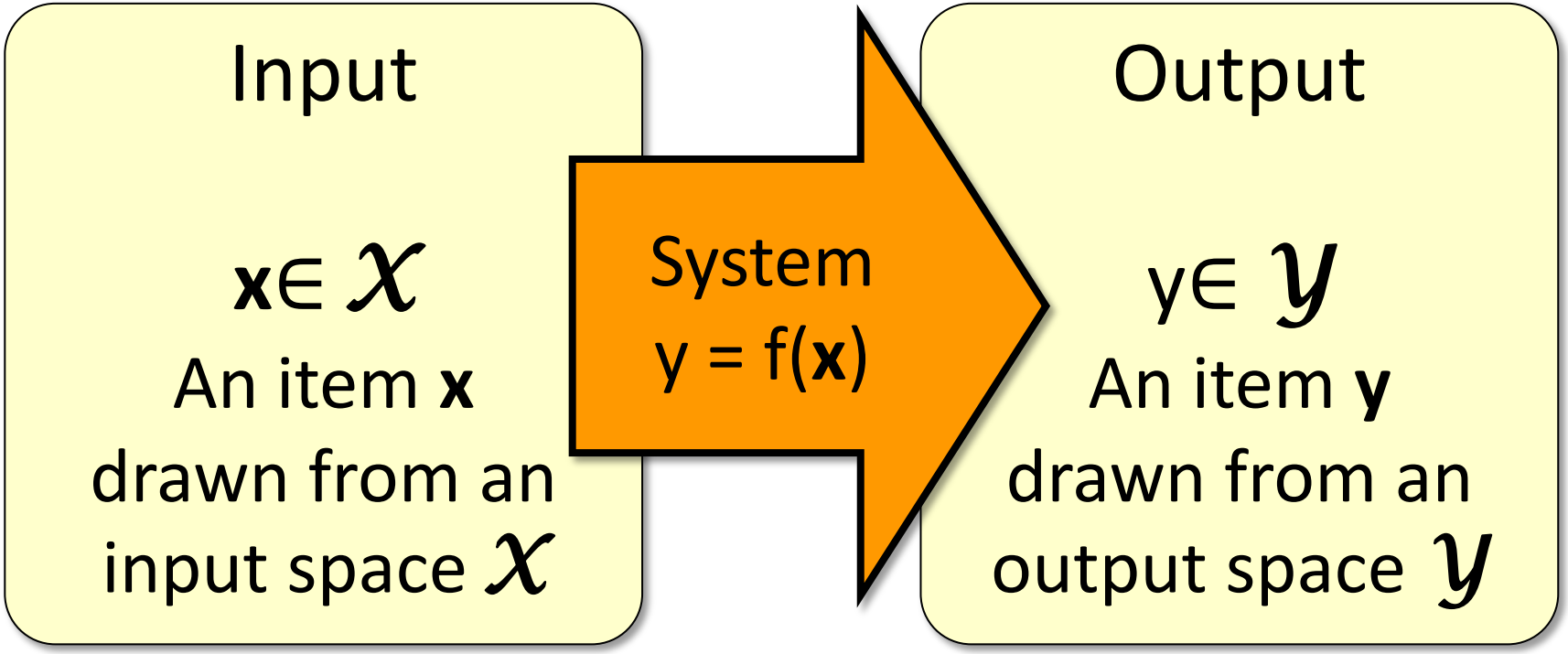
Supervised Learning



- We consider systems that apply a function f() to input items **x** and return an output **y** = f(**x**).

Slide adopted from Dan Roth

Supervised Learning



- In (supervised) machine learning, we deal with systems whose f(**x**) is learned from examples.

Slide adopted from Dan Roth

Classification for WSD

- A **classifier** that can be used to assign sense to new unseen examples
- **Classification** is the problem of identifying to which of a set of categories a particular observation belongs:
  - The badges game is a classification task
  - WSD is a classification task
- A **classifier** can be built by hand – How?
  - But it is not likely to do well
    - Why?
- Machine learning algorithms can be used to build a classifier

Classification 101 (Supervised learning)

Example: Word Sense Disambiguation

Task: Build a model that **learns to classify** the occurrences of the word *bass* using training data: **{fish, music}**

```
(1) The bass guitar is a musical instrument.
(2) Largemouth bass is a warm water fish.
```

How?

- **Labeled training data:** (X, Y): $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- **Feature function:** *Φ(X),* e.g. context features (words and word combinations)
- **Training: use a supervised machine learning algorithm to learn the typical contexts** that correspond to each of the senses: h : X → Y.
- Now we have a **function** that can be applied to classify new occurrences

Feature extraction

- (1) The bass guitar is a musical instrument.
- (2) Largemouth bass is a warm water fish.

What information can we use to determine the sense of the word?

- **Bag-of-words** – unordered set of words in context
- **Collocations** – a word or phrase in a specific position with respect to the ambiguous word
- **Parts-of-speech**

Selecting good features is important

Feature extraction

- (1) The **bass** guitar is a musical instrument.
- (2) Largemouth **bass** is a warm water fish.

What information can we use to determine the sense of the word?

Collocations:

Feature vector: [w_{i-2}, POS_{i-2}, w_{i-1}, POS_{i-1}, w_{i+1}, POS_{i+1}, w_{i+2}, POS_{i+2}]

Feature vectors for the above examples:

(1) [-, -, the, DT, guitar, NN, is, VB]

(2) [-, -, largemouth, ADJ, is, VB, a, DT]

Feature extraction

- (1) The bass guitar is a musical instrument.
- (2) Largemouth bass is a warm water fish.

What information can we use to determine the sense of the word?

Bag-of-words (vector of features): each binary feature indicates whether a vocabulary word occurs in the context; size of the vector: number of words in vocabulary:

Example: Vocabulary: [fishing, big, sound, player, fly, rod, pound, playing, guitar, fish, water, instrument]

Feature vectors for the above examples:

(1) [0,0,0,0,0,0,0,0,1,0,0,1]

(2) [0,0,0,0,0,0,0,0,0,0,1,1,0]

Word sense classifier

- Given training data where each training example is represented with a vector of features, we can use a supervised machine learning algorithm to train **a word sense classifier**
- **Supervised machine learning algorithms:**
  - **Naïve Bayes**
  - **Maximum entropy**
  - **Support vector machines**
  - **Etc.**

The Naïve Bayes Classifier

Let $C=\{C_1, C_2,\ldots, C_k\}$ represent the set of classes.
For example, C={fish,music}

Let $x_1 \ldots x_n$ represent a feature vector
Naïve Bayes is a conditional probability model that conditions each possible class on the feature vector:

$$p(C_k \mid x_1 \ldots x_n) = \frac{p(C_k, x)}{p(x)}$$

We can ignore the denominator

The Naïve Bayes Classifier

$$p(C_k \mid x_1 \ldots x_n) = \frac{p(C_k, x)}{p(x)}$$

- Using the chain rule, we can re-write the denominator:

$$\begin{aligned} p(C_k, x) &= p(C_k, x_1, x_2 \ldots x_n) \\ &= p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k, x_1)\, p(x_3 \mid C_k, x_1, x_2) \ldots p(x_n \mid C_k, x_1, x_2 \ldots x_{n-1}) \end{aligned}$$

The Naïve Bayes Classifier

- **Making Naïve Bayes assumptions**:
  - Each feature is conditionally independent of other features given class C

$$p(C_k, x) = p(C_k, x_1, x_2 \ldots x_n)$$
$$= p(C_k)\,p(x_1 \mid C_k)\,p(x_2 \mid C_k)\,p(x_3 \mid C_k) \ldots p(x_n \mid C_k)$$

The Naïve Bayes classifier

The class is chosen that maximizes the following:

$$y^* = \arg\max_{k \in \{1,2,\ldots K\}} p(C_k) \prod_{i=1}^{n} p(x_i | C_k)$$

Training NB

- Training the NB classifier consists of estimating two types of parameters
  - The prior probability for each class
  - The individual features probabilities

Parameter estimation for NB

- Prior probability:

$$p(C_i) = \frac{count(C_i)}{\sum_{k=1}^{K} count(C_k)}$$

- Context features:

$$p(f_i \mid C_k) = \frac{count(f_i, C_k)}{count(C_k)}$$

Text Classification and Naïve Bayes

Is this spam?

Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

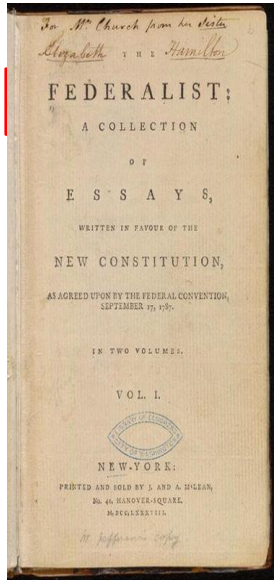
Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

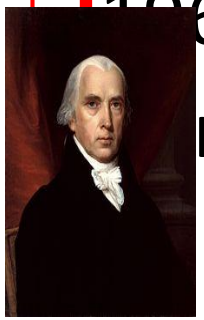
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

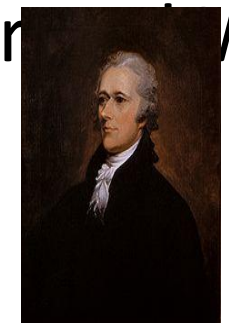
Who wrote which Federalist pa



- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller[illegible]Wallace [illegible]ng Bayesian methods



James Madison



Alexander Hamilton

Male or female author?

1. 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…
2. 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

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.

What is the subject of this article?

MEDLINE Article

Brain and Cognition

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Text Classification

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

Text Classification: definition

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

- *Output*: a predicted class $c \in C$

Classification Methods: Supervised Machine Learning

- *Input:*
  - a document *d*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma:d \rightarrow c$

Text Classification and Naïve Bayes

The Task of Text Classification

Text Classification and Naïve Bayes

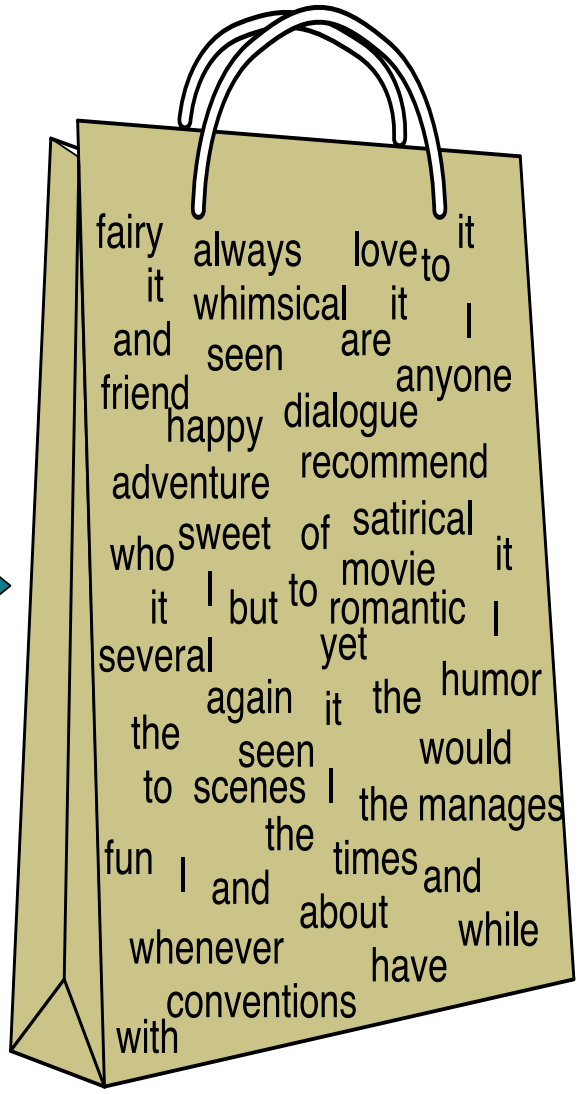
Naïve Bayes (I)

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
...	...

The bag of words representation

$\gamma($

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

$)=c$

Text Classification and Naïve Bayes

Naïve Bayes (I)

Text Classification and Naïve Bayes

Formalizing the Naïve Bayes Classifier

Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Naïve Bayes Classifier (I)

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

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

Naïve Bayes Classifier (II)

$$c_{MAP} = \underset{c \hat{I} C}{\operatorname{argmax}} P(d \mid c) P(c)$$

Document d represented as features x1..xn

$$= \underset{c \in C}{\operatorname{argmax}} \; P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Multinomial Naïve Bayes Independence Assumptions

$$P(x_1, x_2, \ldots, x_n \mid c)$$

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

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Multinomial Naïve Bayes Classifier

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

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

Applying Multinomial Naive Bayes Classifiers to Text Classification

positions ← all word positions in test document

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

Text Classification and Naïve Bayes

Formalizing the Naïve Bayes Classifier

Text Classification and Naïve Bayes

Naïve Bayes: Learning

Learning the Multinomial Naïve Bayes Model

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

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

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

Parameter estimation

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

fraction of times word w_i appears among all words in documents of topic c_j

- Create mega-document for topic *j* by concatenating all docs in this topic
  - Use frequency of *w* in mega-document

Problem with Maximum Likelihood

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

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

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

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

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

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

	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:

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

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

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

P(c|d5) $\propto$ 3/4 * (3/7)^3 * 1/14 * 1/14
≈ 0.0003

P(j|d5) $\propto$ 1/4 * (2/9)^3 * 2/9 * 2/9
≈ 0.0001

Text Classification and Naïve Bayes

Formalizing the Naïve Bayes Classifier