

# Probability Review and Naïve Bayes

Instructor: Alan Ritter

Some slides adapted from Dan Jurfasky and Brendan O'Connor

# What is Probability?

- “The probability the coin will land heads is 0.5”
  - Q: what does this mean?
- 2 Interpretations:
  - Frequentist (Repeated trials)
    - If we flip the coin many times...
  - Bayesian
    - We believe there is equal chance of heads/tails
    - Advantage: events that do not have long term frequencies



Q: What is the probability the polar ice caps will melt by 2050?

# Probability Review

$$\sum_x P(X = x) = \text{?}$$

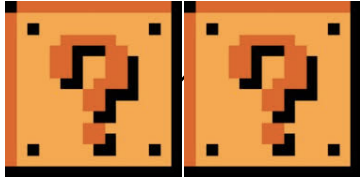
Conditional  
Probability

$$\frac{P(A, B)}{P(B)} = \text{?}$$


Chain Rule

$$P(A|B)P(B) = \text{?}$$


# Probability Review

$$\sum_x P(X = x, Y) =$$


Disjunction / Union:

$$P(A \vee B) =$$


Negation:  $P(\neg A) =$



# Bayes Rule

Hypothesis  
(Unknown)

Data  
(Observed  
Evidence)

$$P(D|H)$$

Generative Model of How Hypothesis Causes Data

$H$

$D$

Bayesian Inference

$$P(H|D)$$

Bayes Rule tells us how to flip the conditional  
Reason about effects to causes

Useful if you assume a generative model for your data

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

# Bayes Rule

Bayes Rule tells us how to flip the conditional  
Reason about effects to causes  
Useful if you assume a generative model for your data

The diagram illustrates the components of Bayes' Rule. The formula is  $P(H|D) = \frac{P(D|H)P(H)}{P(D)}$ . Arrows point from labels to parts of the formula: 'Likelihood' points to  $P(D|H)$ , 'Prior' points to  $P(H)$ , 'Posterior' points to  $P(H|D)$ , and 'Normalizer' points to  $P(D)$ .

Likelihood

Prior

Posterior

Normalizer

$$P(H|D) = \frac{P(D|H)P(H)}{P(D)}$$

# Bayes Rule

Bayes Rule tells us how to flip the conditional  
Reason about effects to causes  
Useful if you assume a generative model for your data

The diagram illustrates the components of Bayes' Rule. The formula is  $P(H|D) = \frac{P(D|H)P(H)}{\sum_h P(D|H)P(H)}$ . Arrows point from labels to parts of the formula: 'Likelihood' points to  $P(D|H)$  in the numerator; 'Prior' points to  $P(H)$  in the numerator; 'Posterior' points to  $P(H|D)$  on the left; and 'Normalizer' points to the denominator  $\sum_h P(D|H)P(H)$ .

Likelihood

Prior

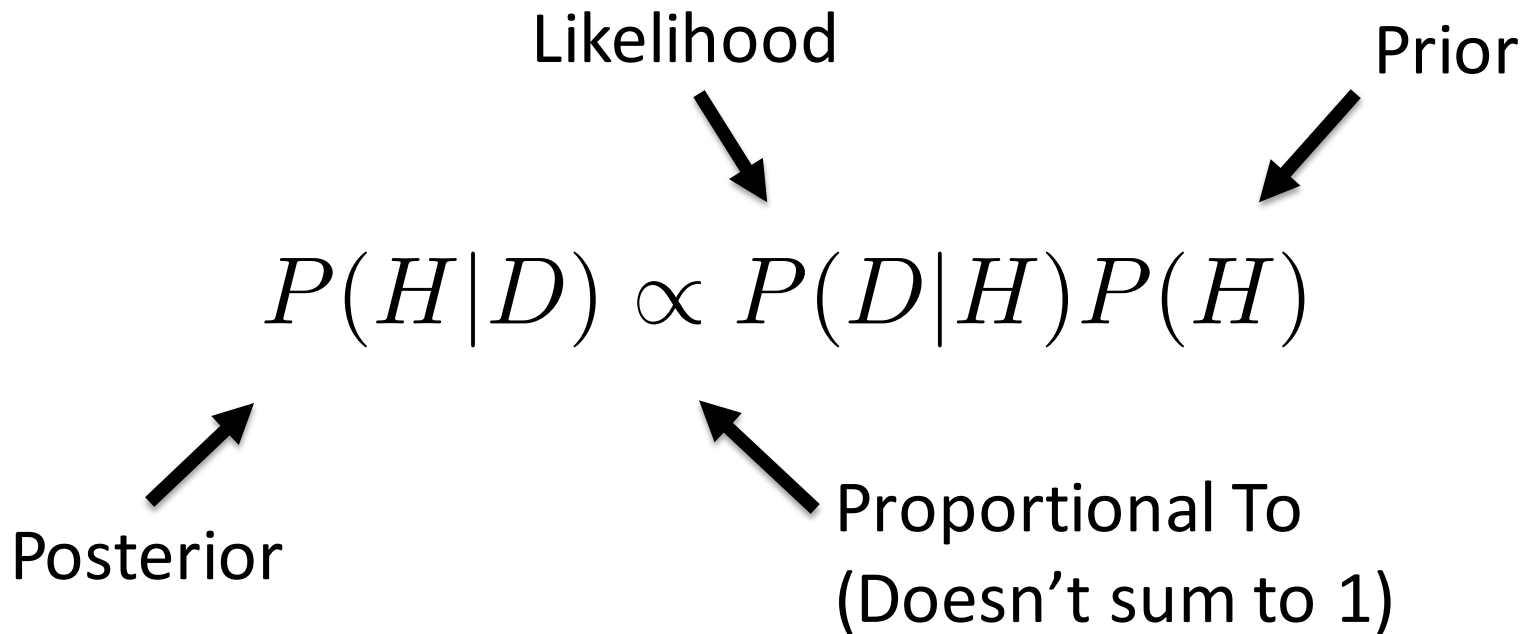
Posterior

Normalizer

$$P(H|D) = \frac{P(D|H)P(H)}{\sum_h P(D|H)P(H)}$$

# Bayes Rule

Bayes Rule tells us how to flip the conditional  
Reason about effects to causes  
Useful if you assume a generative model for your data





# Bayes Rule Example

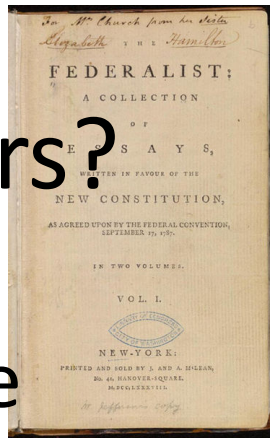
- There is a disease that affects a tiny fraction of the population (0.01%)
- Symptoms include a headache and stiff neck
  - 99% of patients with the disease have these symptoms
- 1% of the general population has these symptoms
- Q: assume you have the symptom, what is your probability of having the disease?

# Text Classification

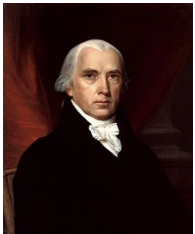
# Is this Spam?



# Who wrote which Federalist papers?



- 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 and Wallace using Bayesian methods



James Madison



Alexander Hamilton

# What is the subject of this article?

## MEDLINE Article



## MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

# 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.

# Text Classification: definition

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



# Classification Methods: Supervised Machine Learning

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

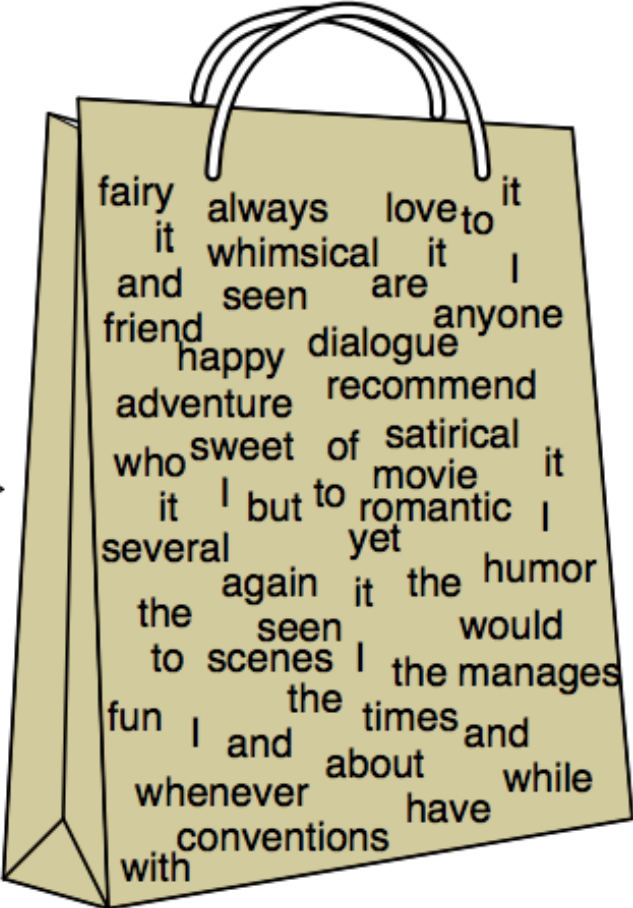
# Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Support-vector machines
  - k-Nearest Neighbors
  - ...

# Naïve Bayes Intuition

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

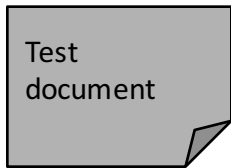
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy 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
...	...

# Bag of words for document classification

?



parser  
language  
label  
translation  
...

Machine  
Learning

learning  
training  
algorithm  
shrinkage  
network...

NLP

parser  
tag  
training  
translation  
language...

Garbage  
Collection

garbage  
collection  
memory  
optimization  
region...

Planning

planning  
temporal  
reasoning  
plan  
language...

GUI

...

# Bayes' Rule Applied to Documents and Classes

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

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

# 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

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

Dropping the denominator

# Naïve Bayes Classifier (II)

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

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



# Naïve Bayes Classifier (IV)

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

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

Could only be estimated if a very, very large number of training examples was available.

How often does this class occur?

We can just count the relative frequencies in a corpus

# Multinomial Naïve Bayes Independence Assumptions

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

- **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  $c$ .

# Multinomial Naïve Bayes Classifier

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

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

# Applying Multinomial Naive Bayes Classifiers to Text Classification

positions  $\leftarrow$  all word positions in test document

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

## Learning the Multinomial Naïve Bayes Model

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

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

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

# Parameter estimation

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

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!

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

# Laplace (add-1) smoothing for Naïve Bayes

$$\begin{aligned}\hat{P}(w_i | c) &= \frac{\textit{count}(w_i, c)}{\sum_{w \in V} (\textit{count}(w, c))} \\ &= \frac{\textit{count}(w_i, c) + 1}{\left( \sum_{w \in V} \textit{count}(w, c) \right) + |V|}\end{aligned}$$



# Multinomial Naïve Bayes: Learning

- Calculate  $P(c_j)$  terms

- For each  $c_j$  in  $C$  do

- $docs_j \leftarrow$  all docs with class  $= c_j$

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

# Multinomial Naïve Bayes: Learning

From training corpus, extract *Vocabulary*

Calculate  $P(w_k | c_j)$  terms

- $Text_j \leftarrow$  single doc containing all *docs*<sub>*j*</sub>
- For each word  $w_k$  in *Vocabulary*  
 $n_k \leftarrow$  # of occurrences of  $w_k$  in  $Text_j$

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

# Exercise