Probability Review and Naïve Bayes

Instructor: Wei Xu

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
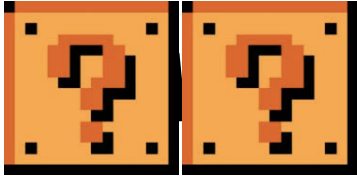   - Advantage: events that do not have long term

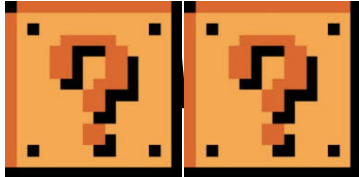E.g. What is the probability the polar ice caps will melt by 2050?
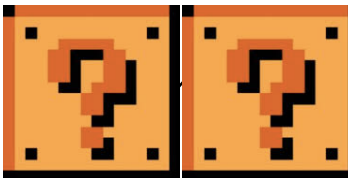
Probability Review

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

Conditional Probability

$$\frac{P(A, B)}{P(B)} = \boxed{?}\,\boxed{?}$$
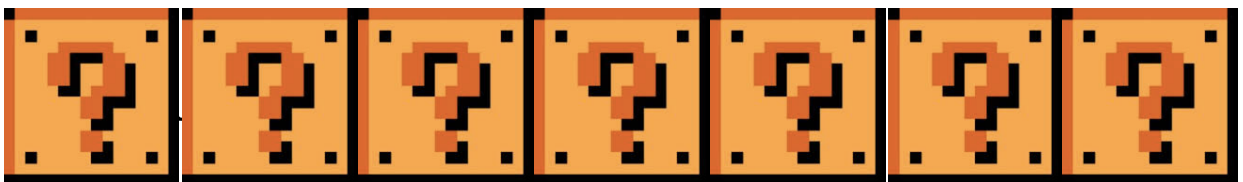
Chain Rule

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

Probability Review

$$\sum_x P(X = x, Y) = \boxed{?}\,\boxed{?}$$

Disjunction / Union:

$$P(A \vee B) = \boxed{?}\,\boxed{?}\,\boxed{?}\,\boxed{?}\,\boxed{?}\,\boxed{?}\,\boxed{?}$$

Negation: $P(\neg A) = \boxed{?}\,\boxed{?}\,\boxed{?}$

Bayes Rule

Hypothesis (Unknown)

Data (Observed Evidence)

$$P(D|H)$$

$$H \xrightarrow{\text{Generative Model of How Hypothesis Causes Data}} D$$

Bayesian Inferece

$$P(H|D)$$
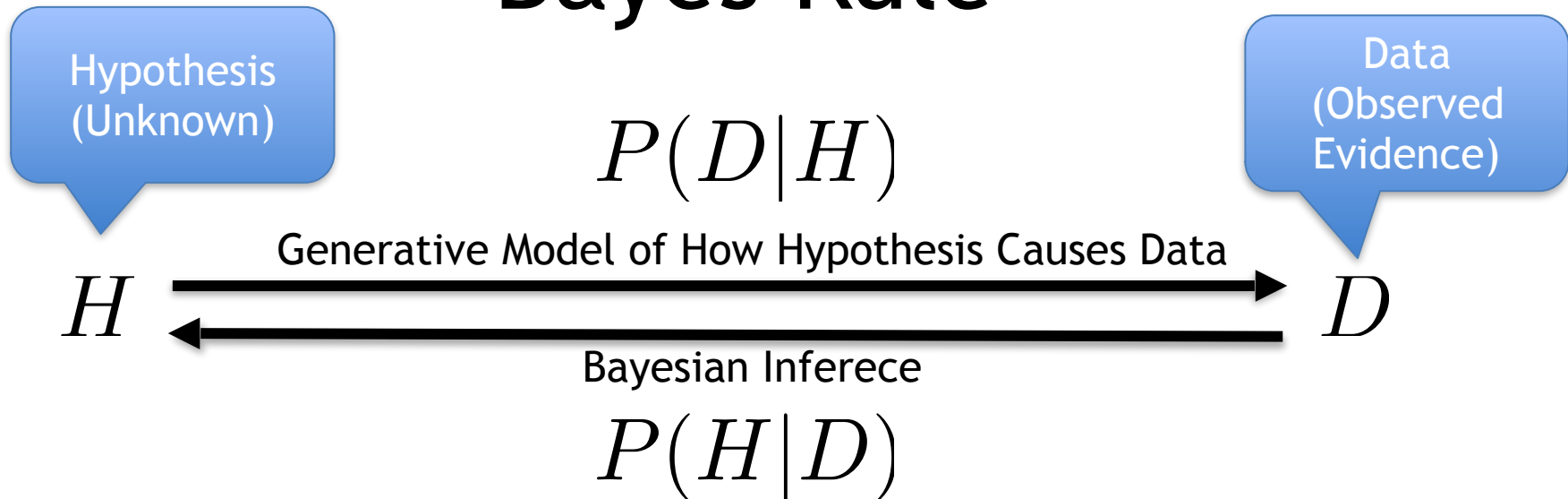
Bayes Rule tells us how to "flip" the conditional probabilities
Reason about effects to causes
Useful if you assume a generative model for your data

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

Bayes Rule

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

Likelihood

Prior

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

Posterior

Normalizer

Bayes Rule

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

Likelihood

Prior

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

Posterior

Normalizer

Bayes Rule

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

Likelihood

Prior

$$P(H|D) \propto P(D|H)P(H)$$

Posterior

Proportional To
(Doesn't sum to 1)

Bayes Rule Example

- There is a disease that affects a tiny fraction of the population (0.001%)
- Symptoms include a headache and stiff neck. 50% of patients with the disease have these symptoms
- 5% of the general population has these symptoms.

Q: Assume you have the symptom, what is your probability of having the disease?

Another Bayes Rule Example

- The well-known OJ Simpson murder trial

Another Bayes Rule Example

- The prosecution presented evidence that Simpson had been violent toward his wife, argued that a pattern of spousal abuse reflected a motive to kill.

- The defense attorney, Alan Dershowitz, argued that:
  - there was only one woman murdered for every 2500 women who were subjected to spousal abuse, and that any history of Simpson being violent toward his wife was irrelevant to the trial.

- In effect, both sides were asking the jury to consider the probability that a man murdered his ex-wife, given that he previously battered her.

What do you think?
Discuss with your neighbors

Another Bayes Rule Example

- The defense attorney, Alan Dershowitz, argued that:
  - there was only one woman murdered for every 1000 women who were subjected to spousal abuse, and that any history of Simpson being violent toward his wife was irrelevant to the trial.
- In 1994, 5000 women were murdered, 1500 by their husband. Assuming a population of 100 million women.
  - P (Murder|¬Guilt) = $3500/100 \times 10^6 \approx 1/30000$

What do you have now?
Discuss with your neighbors

Text Classification

Is this Spam?

Act Fast — Junk

PUBLISHERS CLEARING HOUSE   December 22, 2015 at 11:34 PM

To: Undisclosed recipients:;

Reply-To: william.kelley0@accountant.com

Act Fast

FROM THE DESK OF
DEBORAH HOLLAND
EXECUTIVE VICE PRESIDENT.
PUBLISHERS CLEARING HOUSE

You have been declared the Publishers Clearing House mega winner of $550,000.00 in the recent on going draw held on 20th of December 2015.

For verification Contact William Kelley, Reply and fill in the infomation below ASAP .

ADDRESS........
STATE.............
CITY...............
SEX................
AGE..............
CELLPHONE....

Congratulations once again.

Who wrote which Federalist papers?
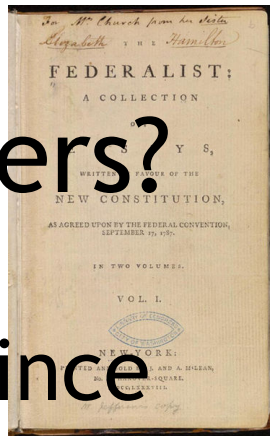
- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.

- Authorship of 12 of the letters in dispute

- 1963: solved by Mosteller and Wallace using Bayesian methods
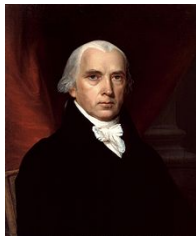
James Madison

Alexander Hamilton

What is the subject of this article?

MEDLINE Article

MeSH Subject Category Hierarchy

- Antogonists 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 $C = \{c_1, c_2, ..., c_J\}$

- *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
- Running time is usually very good and fast
- But, building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

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

- 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!

fairy always love to it whimsical it and seen are I friend anyone happy dialogue adventure recommend who sweet of satirical it it I but to movie it several yet romantic I again it the humor the seen would to scenes I the manages fun I the times and and about while whenever have conventions with
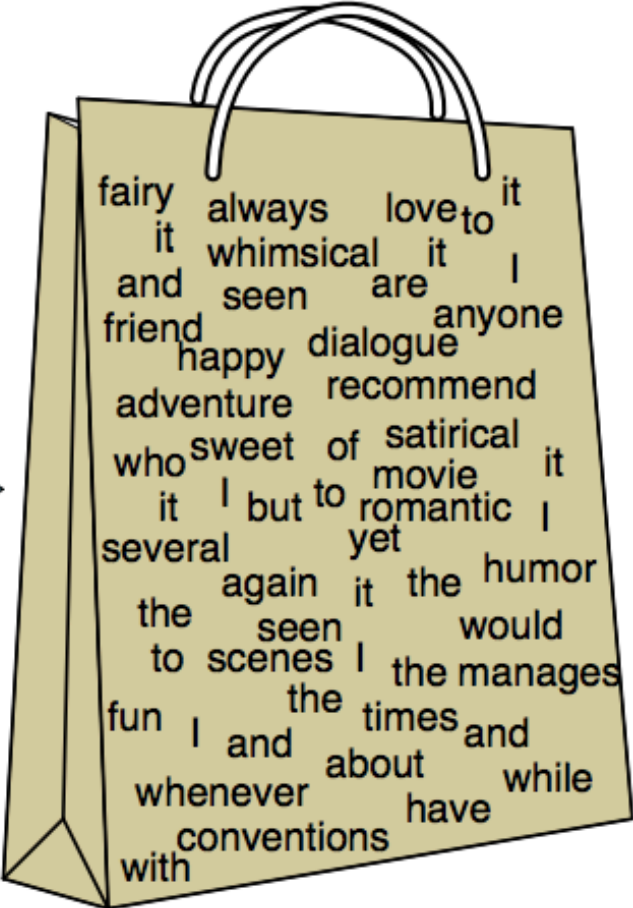
it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Bag of words for document classification

Test document
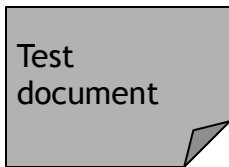
parser
language
label
translation
…

?

Machine Learning	NLP	Garbage Collection	Planning	GUI
learning	parser	garbage	planning	…
training	tag	collection	temporal	
algorithm	training	memory	reasoning	
shrinkage	translation	optimization	plan	
network…	language…	region…	language…	

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)

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

MAP is "maximum a posteriori" = most likely class

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

Bayes Rule

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

Dropping the denominator

Naïve Bayes Classifier (II)

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

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

Naïve Bayes Classifier (IV)

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

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

How often does this class occur?

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

We can just count the relative frequencies in a corpus

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 $P(x_i | c_j)$ are independent given the class c.

Multinomial Naïve Bayes Classifier

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

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

Applying Multinomial Naive Bayes to Text Classification

positions ← all word positions in test document

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

Multinomial Naïve Bayes: Learning

- 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

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

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)*?

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

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

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

Laplace (add-1) smoothing for Naïve Bayes

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c)}{\displaystyle\sum_{w \in V}\left(count(w, c)\right)}$$

$$= \frac{count(w_i, c) + 1}{\left(\displaystyle\sum_{w \in V} count(w, c)\right) + |V|}$$

Multinomial Naïve Bayes: Learning

- Calculate $P(c_j)$ terms
  - For each c_j in C do

  $$docs_j \leftarrow \text{ all docs with class } = c_j$$

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

Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*
- Calculate $P(w_k \mid c_j)$ terms
  - *Text$_j$* ← single doc containing all *docs$_j$*
  - For each word w_k in *Vocabulary*

    n_k ← # of occurrences of w_k in *Text$_j$*

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