Probability and Naïve Bayes

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

$$\sum_x P(X = x) = 1$$

Conditional Probability

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

Chain Rule

$$P(A|B)P(B) = P(A, B)$$

Probability Review

Disjunction / Union: $$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Negation: $$P(\neg A) = 1 - P(A)$$

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

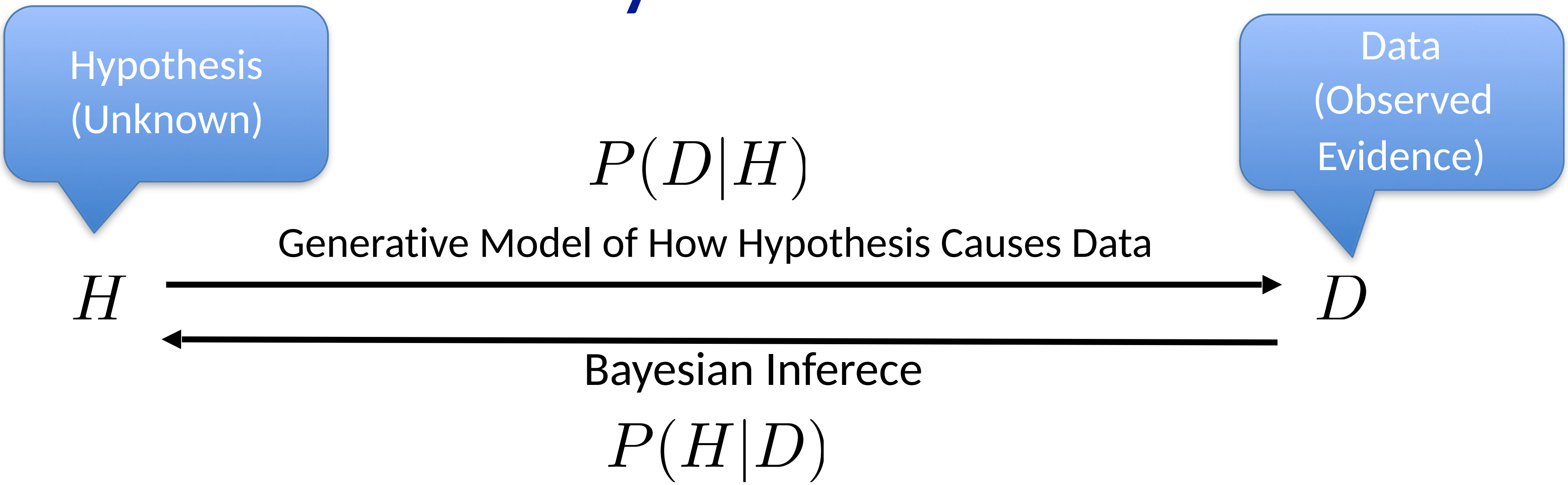
Probability Review

Disjunction / Union: $$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

Negation: $$P(\neg A) = 1 - P(A)$$

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

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

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

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

Likelihood

Prior

$$\underbrace{P(H|D)}_{\text{Posterior}} = \frac{\overbrace{P(D|H)}^{\text{Likelihood}}\,\overbrace{P(H)}^{\text{Prior}}}{\underbrace{P(D)}_{\text{Normalizer}}}$$

Posterior

Normalizer

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

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
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)

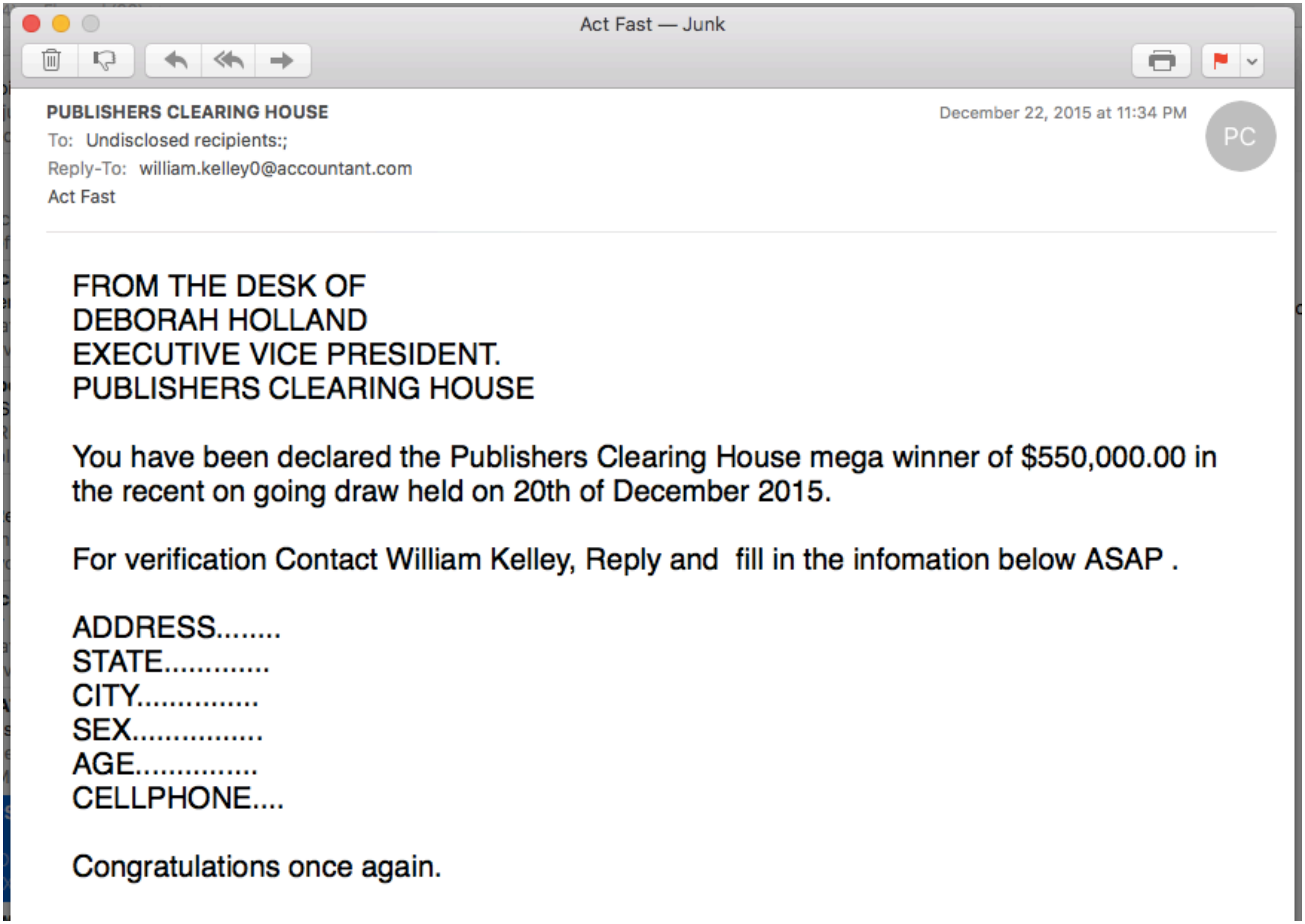
Bayes Rule Example

- There is a disease that affects a tiny fraction of the population (0.01%)
- 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?



Act Fast — Junk

PUBLISHERS CLEARING HOUSE
To: Undisclosed recipients:;
Reply-To: william.kelley0@accountant.com
Act Fast

December 22, 2015 at 11:34 PM

PC

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.

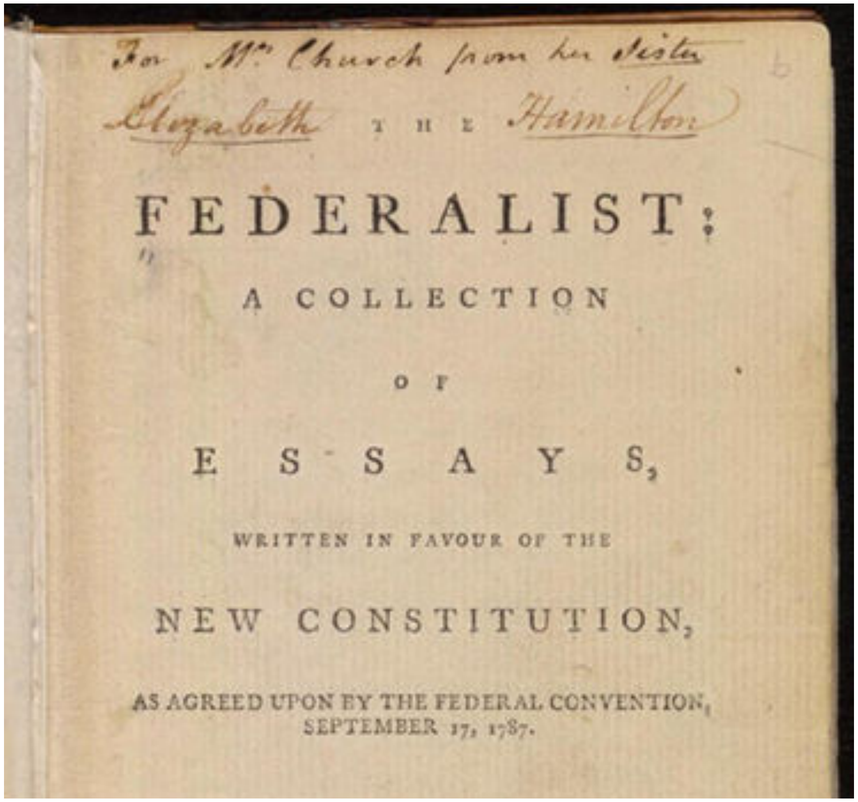
For verification Contact William Kelley, Reply and fill in the infomation below ASAP .

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.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods



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
- …

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 $C = \{c_1, c_2, \ldots, c_J\}$

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

Classification Methods: Supervised Machine Learning

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

  - …

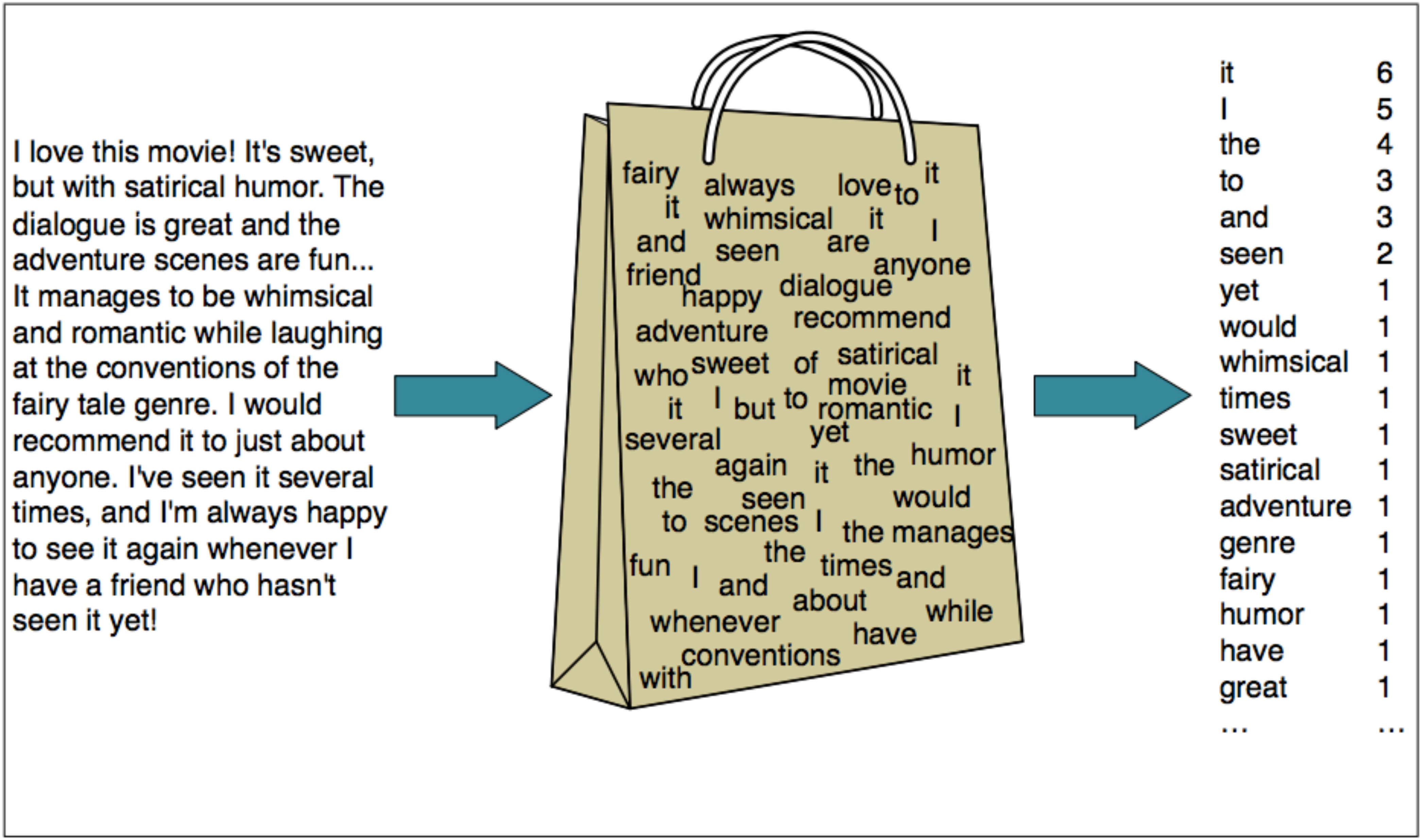
Classification Methods: Supervised Machine Learning

- 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!
fairy always love to it
it whimsical it I
and seen are
friend anyone
happy dialogue
adventure recommend
who sweet of satirical
movie it
it I but to romantic I
several yet
again it the humor
the seen would
to scenes I the manages
fun the times
I 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
... ...

Bayes' Rule Applied to Documents and Classes

For a document d and a class c

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

Bayes' Rule Applied to Documents and Classes

For a document d and a class c

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

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

Naïve Bayes Classifier (I)

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

MAP is "maximum a posteriori" = most likely class

Naïve Bayes Classifier (I)

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

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

MAP is "maximum a posteriori" = most likely class

Bayes Rule

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)

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

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

Naïve Bayes Classifier (IV)

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

How often does this class occur?

We can just count the relative frequencies in a corpus

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

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(xi|cj) are independent given the class c.

Multinomial Naïve Bayes Classifier

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

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

Applying Multinomial Naive Bayes Classifiers to Text Classification

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

positions $\leftarrow$ all word positions in test document

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

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

fraction of word in the full vocabulary that appered in topic 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

fraction of word in the full vocabulary that appered in topic 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)***?

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

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

Laplace (add-1) smoothing for Naïve Bayes

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

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

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
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

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

Exercise

Multinomial Naïve Bayes: Learning

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

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

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

- From training corpus, extract *Vocabulary*
- Calculate $P(w_k \mid c_j)$ terms
  - $Text_j \leftarrow$ single doc containing all $docs_j$
  - For each word w_k in *Vocabulary*

    $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

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

Solution

	Category	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprise and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no originality

$$P(-) = \frac{3}{5} \qquad P(+) = \frac{2}{5}$$

	Category	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprise and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no originality

$$P(\text{“predictable”}|-) = \frac{1+1}{14+20} \qquad P(\text{“predictable”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“with”}|-) = \frac{0+1}{14+20} \qquad P(\text{“with”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“no”}|-) = \frac{1+1}{14+20} \qquad P(\text{“no”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“originality”}|-) = \frac{0+1}{14+20} \qquad P(\text{“originality”}|+) = \frac{0+1}{9+20}$$

	Category	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprise and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no originality

$$P(S|-)P(-) = \frac{3}{5} \times \frac{2 \times 1 \times 2 \times 1}{34^4} = 1.8 \times 10^{-6}$$

$$P(S|+)P(+) = \frac{2}{5} \times \frac{1 \times 1 \times 1 \times 1}{29^4} = 5.7 \times 10^{-7}$$

The model thus predicts the class *negative* for the test sentence.