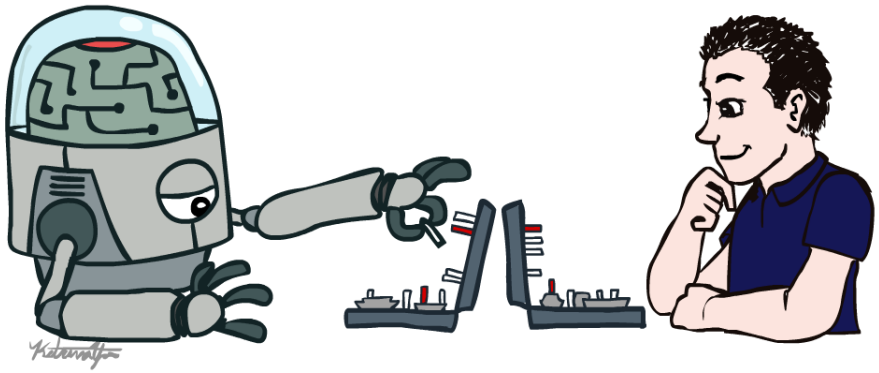
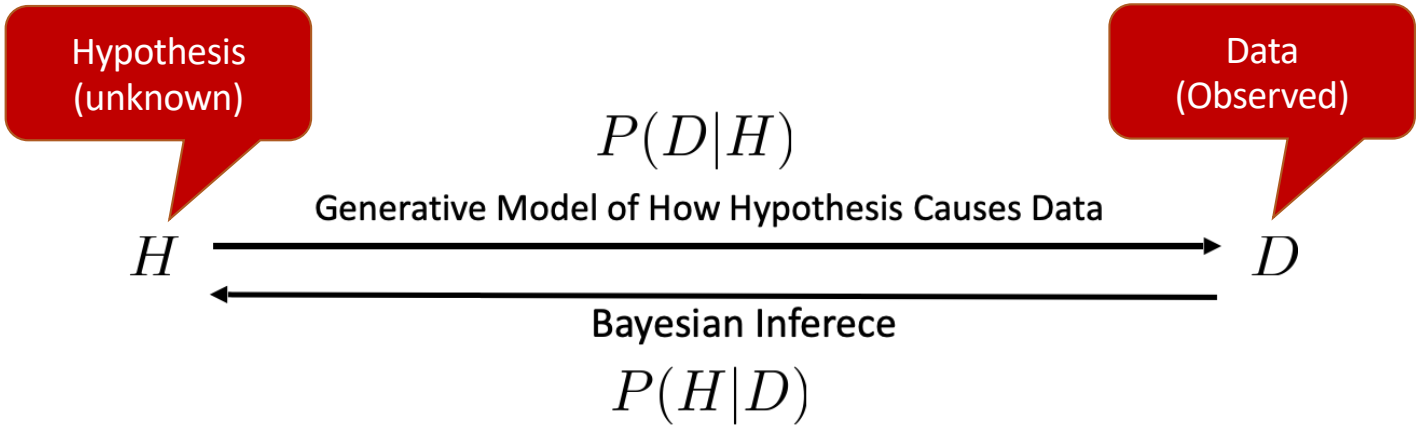
CSE 3521:
Introduction to Artificial Intelligence



[Many slides are adapted from the UC Berkeley. CS188 Intro to AI at UC Berkeley and previous CSE 3521 course at OSU.]

THE OHIO STATE UNIVERSITY

Bayes Rules

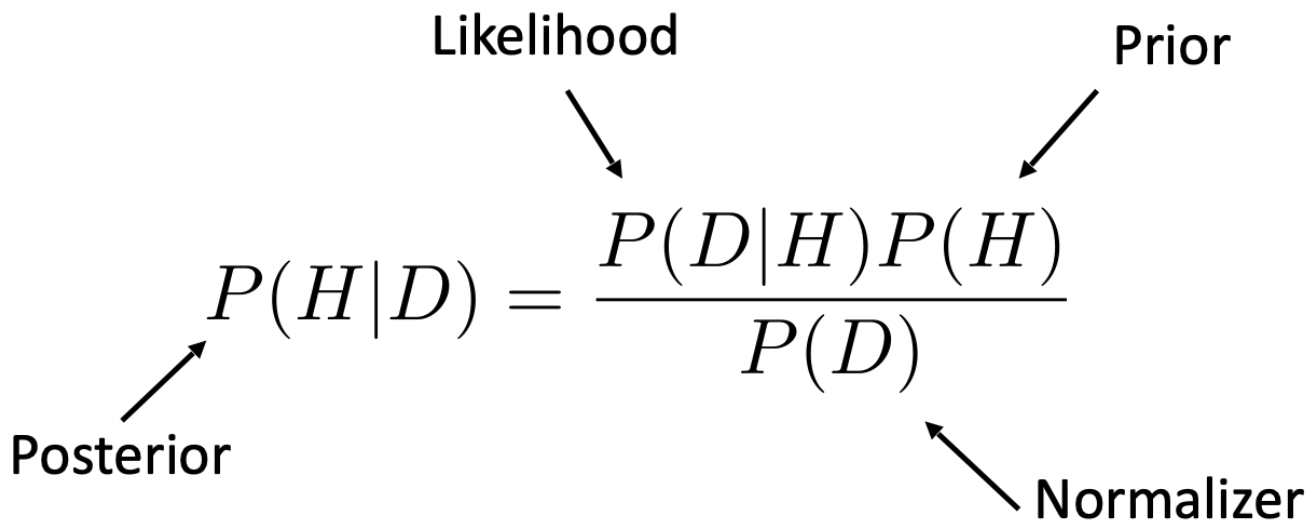


Bayes Rule tells us how to flip the conditional
Reason about effects to causes
Useful if you assume a generative model for your data

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

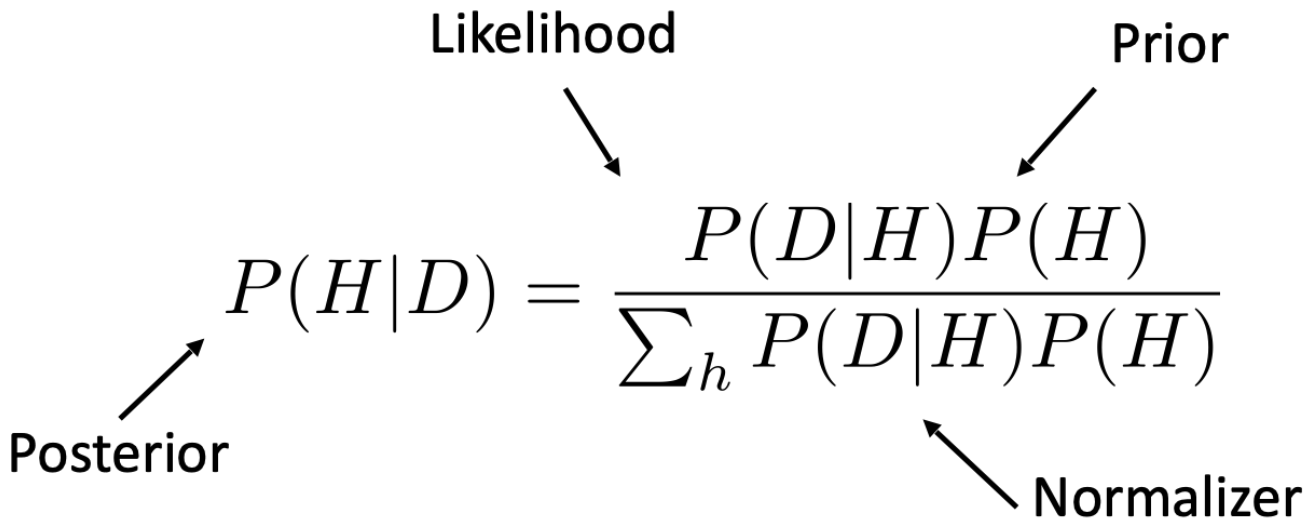
Bayes Rules

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



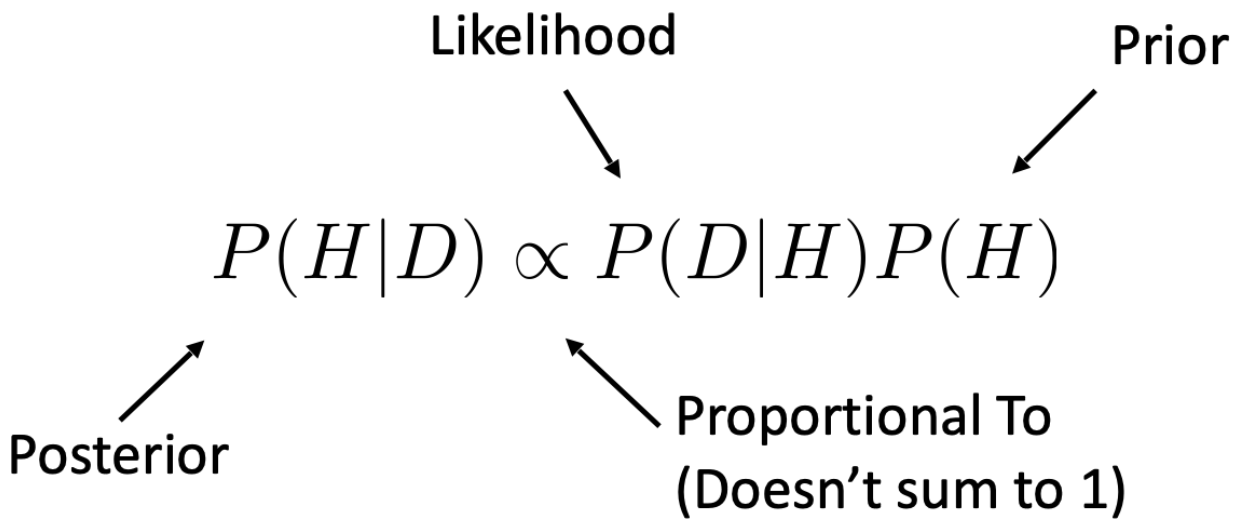
Bayes Rules

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 Rules

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 Rules

- 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

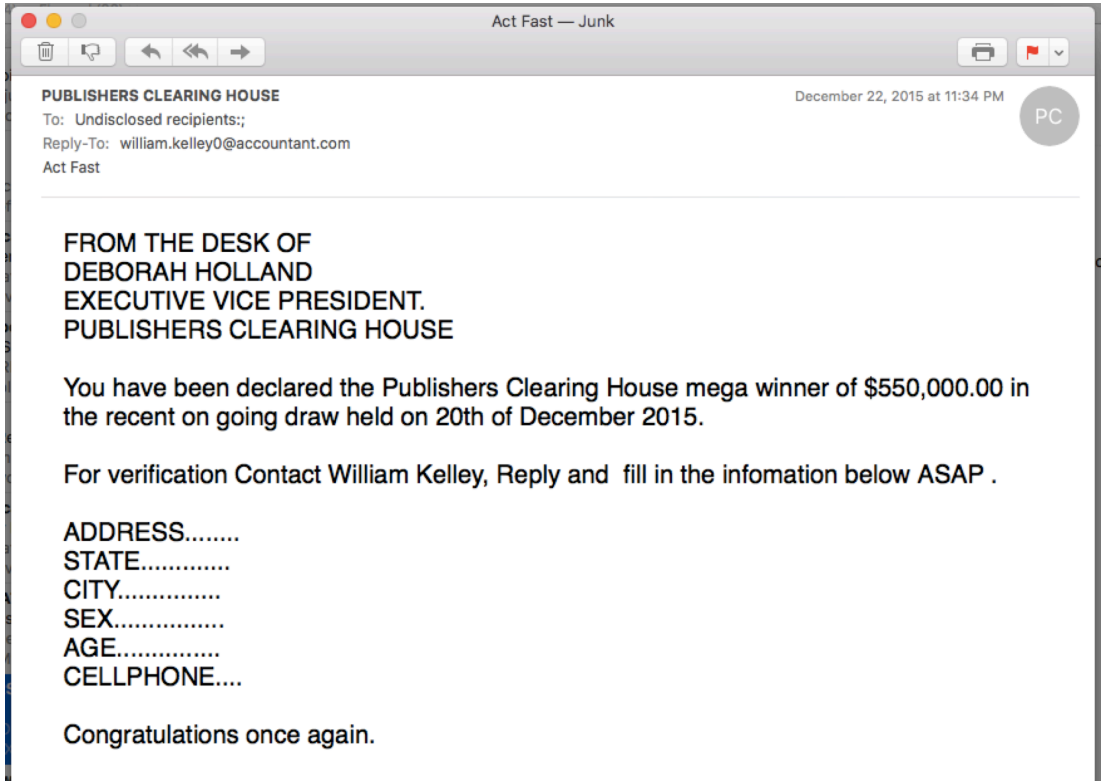
Q: assume you have the symptom, what is your probability of having the disease?

Bayes Rules Examples

- 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

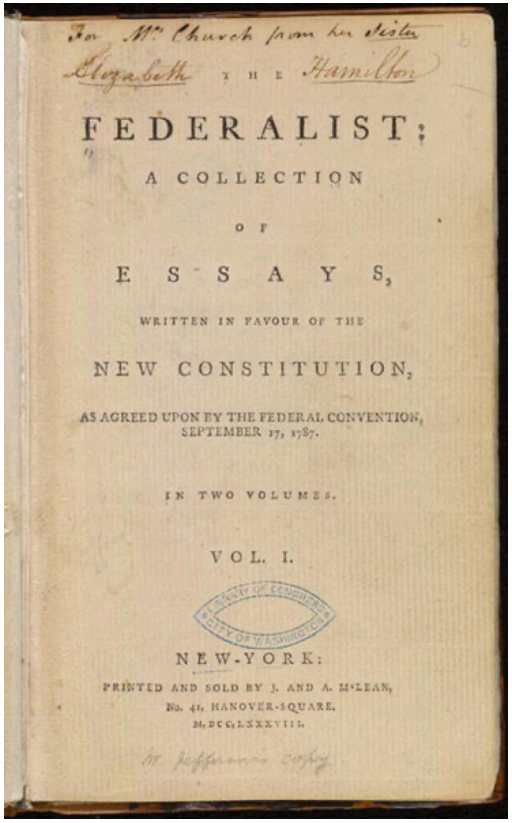
Q: assume you have the symptom, what is your probability of having the disease?

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
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 this paper?

- 1787-88: 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

MeSH Subject Category Hierarchy

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

MEDLINE Article

Brain and Cognition

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

Positive or Negative 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.

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 Method: *Rules*

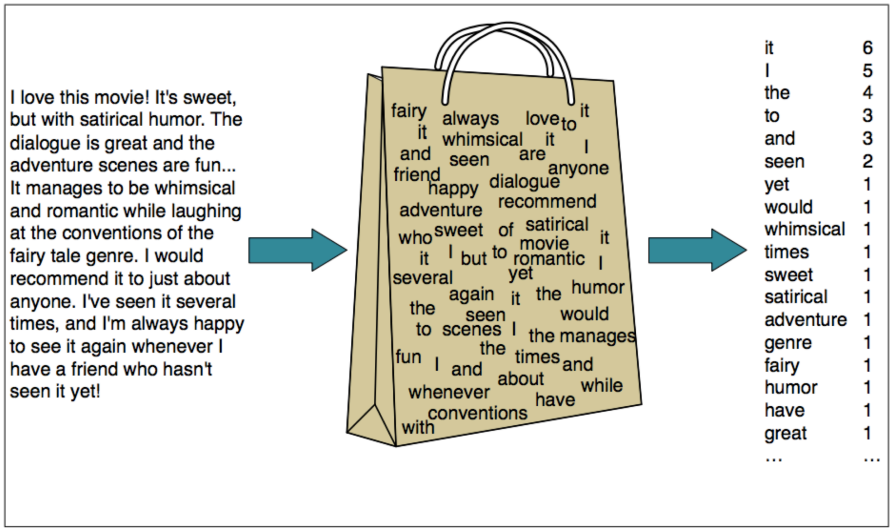
- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Method: *Supervised 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 $\gamma:d \rightarrow c$

Naïve Bayes

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



Bayes Rule Applied to Documents

For a document d and a class c

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

Bayes Rule Applied to Documents

For a document d and a class c

Posterior $\longrightarrow P(c|d) = \dfrac{P(d|c)P(c)}{P(d)}$

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

Naïve Bayes Classifier

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

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

Naïve Bayes Classifier

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

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.

Naïve Bayes Classifier: Independence Assumption

- **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_j.

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

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

Learning: Naïve Bayes Classifier

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

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

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

Learning: Naïve Bayes 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 appeared in topic

Learning: Naïve Bayes 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 appeared in topic

- What if we have seen no training documents with the word ***fantastic*** in the topic **positive**?
- Zero probabilities cannot be conditioned away

Smoothing: Naïve Bayes Parameter Estimation

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

$$\hat{P}(w_i|c) = \frac{count(w_i, c) + \alpha}{\sum_{w \in V} count(w_i, c) + \alpha|V|}$$

Laplace Smoothing: Naïve Bayes Parameter Estimation

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

$$\hat{P}(w_i|c) = \frac{count(w_i, c) + \alpha}{\sum_{w \in V} count(w_i, c) + \alpha|V|}$$

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

Naïve Bayes Parameter Learning: Step

- 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
  - $\text{Text}_j \leftarrow$ single doc containing all sentences from class = c_j
  - For each word w_k in *Vocabulary*

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

    $n \leftarrow$ # of words in class $Text_j$

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

In Class Exercise

What will be category of the test data?

	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

What will be category of the test data?

	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

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

What will be category of the test data?

	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

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

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

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

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

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

What will be category of the test data?

	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

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

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

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

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

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

What will be category of the test data?

	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

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

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

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

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

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

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

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

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