CLASSIFYING DOCUMENTS USING A NAIVE BAYES CLASSIFIER

Probability

I have a coin which is biased towards heads.

Probability of heads = 0.75

(1) QUESTION: What's the probability of getting 9 heads when the coin is tossed 10 times?

↑ This is an easy question

Statistical Inference

I have a coin which I've tossed 10 times and got 9 heads.

② QUESTION: What is the bias of the coin?

↑

This is a hard question!

And it's critical to all of Science.

Let's answer this hard question

Hypothesis The coin's bias towards heads is 0.9.

Evidence / Data Toss the coin 10 times and observe 9 heads in total.

But how to calculate this?

Answer:

$$\text{Prob}(\underbrace{\text{bias} = 0.90}_{H} \mid \underbrace{9 \text{ heads in } 10 \text{ tosses}}_{E})$$

Symbol for "given that"

The structural difference between

$Prob(H \mid E)$ ← What we want (but hard to get)

vs.

$Prob(E \mid H)$ ← What we can usually get

In general,

$$Prob(H \mid E) \neq Prob(E \mid H)$$

Simple Examples

$$\text{Prob}(\text{animal} \mid \text{human}) = 1$$

$$\text{Prob}(\text{human} \mid \text{animal}) \neq 1$$

$$\text{Prob}(\text{dog barks} \mid \text{burglar enters}) = A$$

$$\text{Prob}(\text{burglar enters} \mid \text{dog barks}) = B$$

For many dogs, $A \gg B$

$\gg$ symbol for "much greater than"

$$P(\text{Win the lottery} \mid \text{have \# 1234567})$$

$$= \frac{1}{\#\text{ tickets issued by the lottery Commission}}$$

$$P(\text{have \# 1234567} \mid \text{Win the lottery})$$

$$= \; ? \quad \text{it's either 0 or 1}$$

Usually, it is easy to calculate

Probability calculation

$P(E|H)$ but difficult to

calculate $P(H|E)$.

Statistical inference calculation

AND

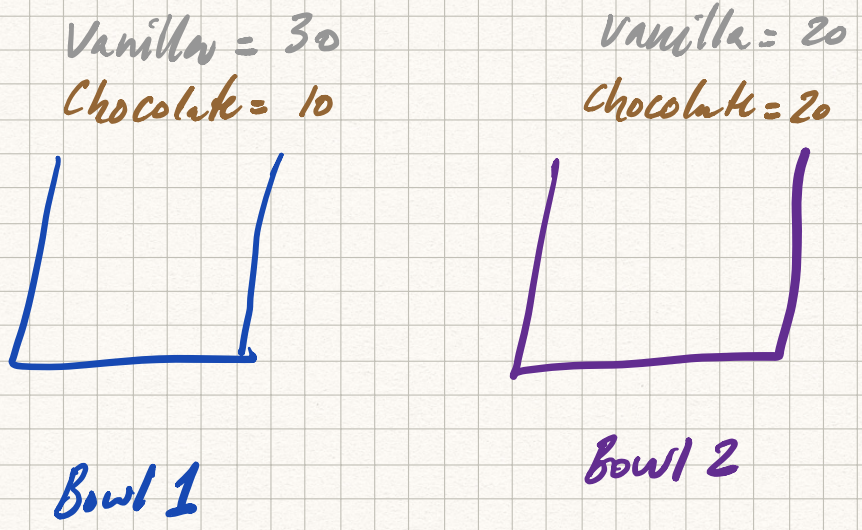
Usually, it's $P(H|E)$ is what we'd most like to know in Science.

Bayes' Rule

$$P(H|E) = \frac{P(E|H)\,P(H)}{P(E)}$$

The Cookie Problem

From Allen B. Downey in *Think Bayes*)



Without looking, pick a cookie from a bowl. Suppose you get a vanilla cookie.

QUESTION: What is the probability that the cookie came from Bowl 1?

$$Prob(Bowl\ 1 \mid vanilla) = ?$$

try

$$Prob(vanilla \mid Bowl\ 1)$$ ← this is easy

$$= \frac{30}{40} = 0.75$$

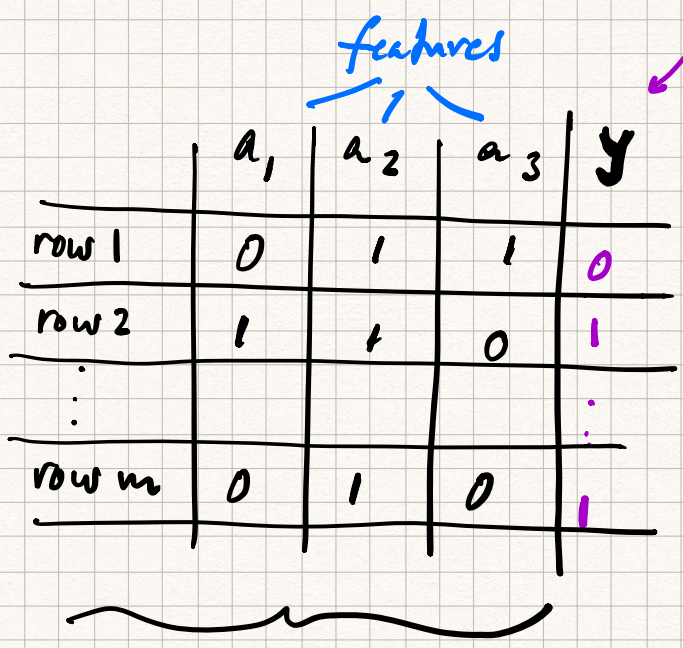
Use Bayes' Theorem

$$Prob(Bowl\ 1 \mid vanilla) = \frac{Prob(vanilla \mid Bowl\ 1) * P(Bowl\ 1)}{Prob(vanilla)}$$

0.75 is $Prob(vanilla \mid Bowl\ 1)$; 0.5 is $P(Bowl\ 1)$

$Prob(vanilla) \rightarrow \frac{50}{80} = 5/8$

OK, what does this have to do spam?



usual logistic regression problem

all features are binary 0 or 1.

Goal: Use this data to learn a classification.

How?

Usual Approach

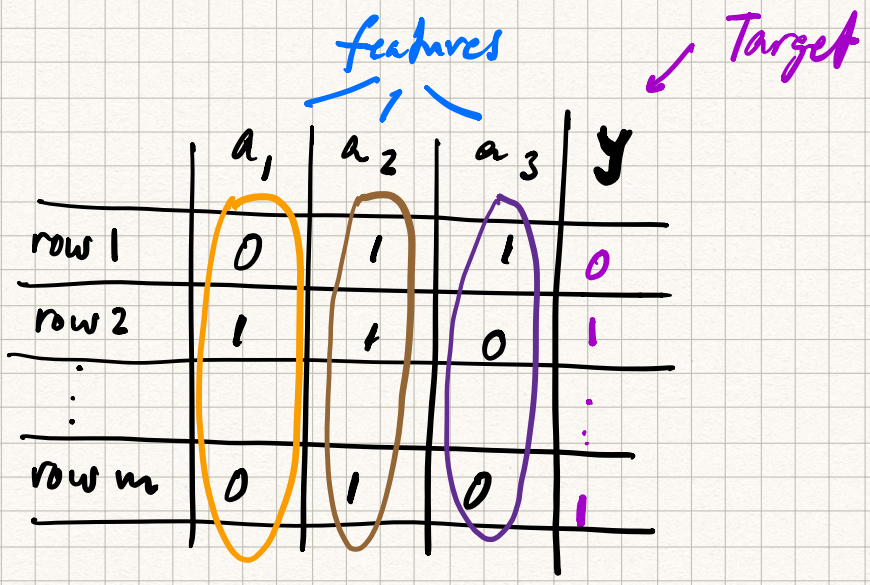
① Traditional logistic regression model:

$$w_0 + a_1 * w_1 + a_2 * w_2 + a_3 * w_3 = \hat{y}$$

② Apply a penalty for each row

③ Find the values of w_0, w_1, w_2, w_3 that minimize the cost (total penalty over the entire dataset).

Another way to approach the problem



Question

$Prob(y=0 \mid a_1=0 \text{ and } a_2=1 \text{ and } a_3=1) = ?$

↓ take advantage of Bayes' theorem

$$= \frac{Prob(a_1=0, a_2=1, a_3=1 \mid y=0) * P(y=0)}{Prob(a_1=0, a_2=1, a_3=1)}$$

Independent Events

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

In words: The probability of A is not affected by the occurrence or non-occurrence of B and vice versa.

$$\frac{Prob(a_1=0, a_2=1, a_3=1 \mid y=0) * P(y=0)}{Prob(a_1=0, a_2=1, a_3=1)}$$ ③

Independence

② $Prob(a_1=0) * Prob(a_2=0) * Prob(a_3=0)$

① $Prob(a_1=0 \mid y=0) * Prob(a_2=1 \mid y=0) * Prob(a_3=1 \mid y=0)$

①, ②, and ③ can be readily calculated from the training data set.

features

Target

	a_1	a_2	a_3	y
row 1	0	1	1	0
row 2	1	1	0	1
⋮				⋮
row m	0	1	0	1

$$P(y=0) = \frac{\#\text{ of } y=0}{\#\text{ rows of data}}$$

$$\text{Prob}(a_1=0) = \frac{\#\text{ of } a_1=0}{\#\text{ rows of data}}$$

Similar expressions for $y=1$

$$\text{Prob}(a_1=0 \mid y=0) = \frac{k}{s}$$

Pick all rows where $y=0$

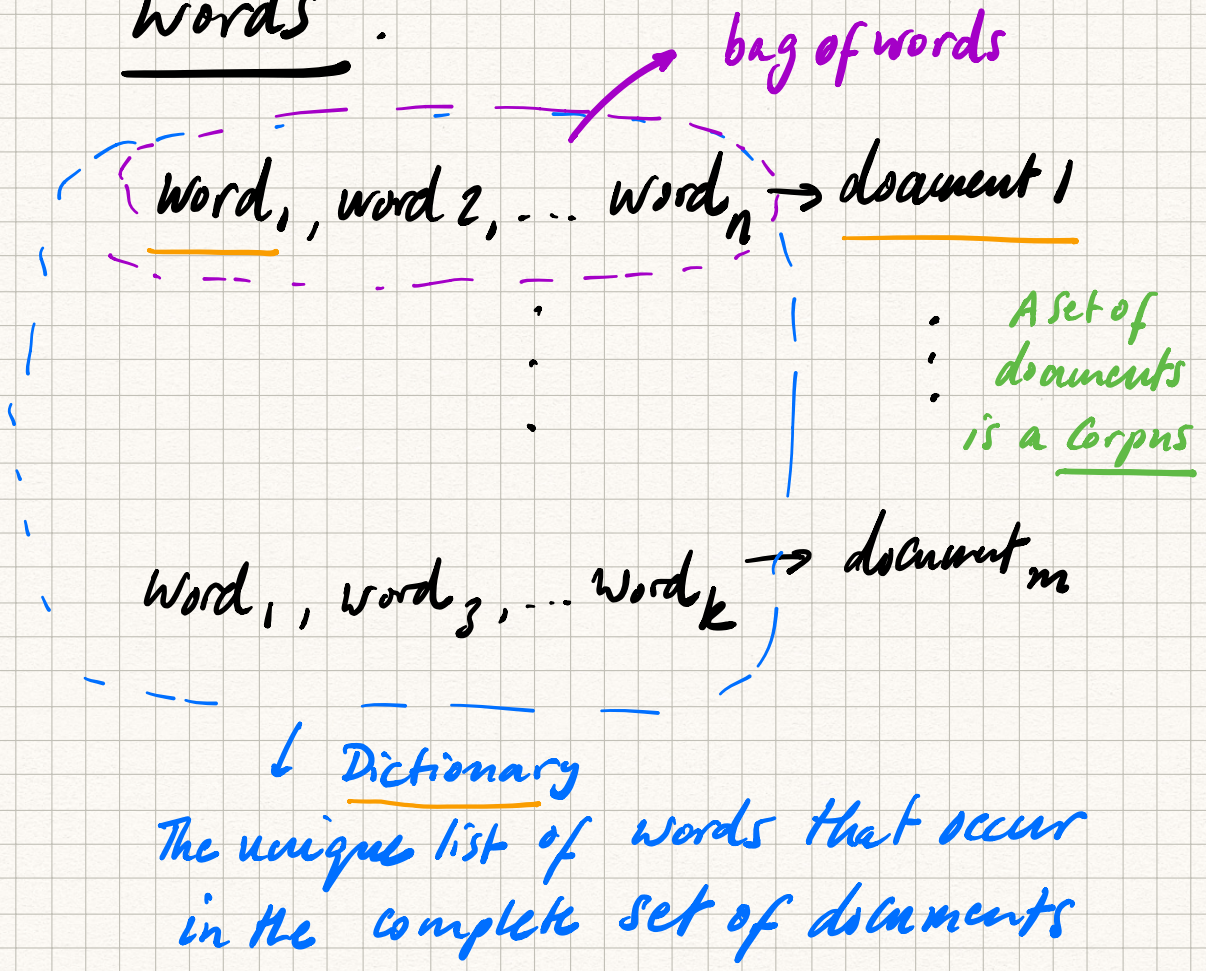
Count the number of these rows ($=s$)

In these rows, count the # of $a_1=0$. ($=k$)

Now we have all we need to use Bayes' rule and calculate the probability of spam or ham.

Where do the $a_1 = 0$, $a_2 = 1$, etc. come from?

They are actually ways to represent **words**.



A simple way to represent a document

= Bag of words

Document 1

"The quick fox jumped over the hen"

↓

Document 1

bag of words

fox
hen
jumped
over
quick
the
the
over

Structure of the sentence is lost.

Word order is not preserved.

Document 2

" Free! Don't Miss Out!! "

↓

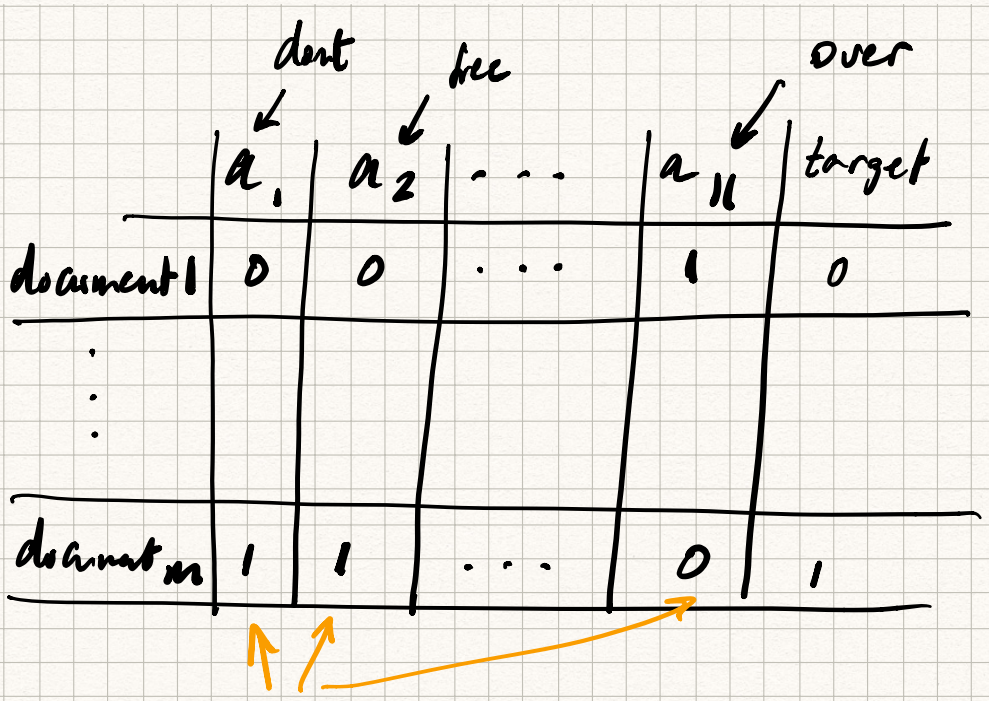
Document 2
bag of words

dont
free
miss
out

} Punctuation ignored.
Capitalization ignored.

Dictionary = { dont, fox, free, hen, ...,
out, over }

There are 11 unique words in the dictionary. These are the **attributes** of the dataset.



The Naive Bayes model is <u>trained</u> on this dataset of documents that are "<u>vectorized</u>" or <u>embedded</u> using the bag of words representation.

What does the Naive Bayes model "learn"?

- $P(a_1 = 0 \mid y = 0)$, $P(a_1 = 1 \mid y = 0)$
  $P(a_1 = 0 \mid y = 1)$, $P(a_1 = 1 \mid y = 1)$
  } for all attributes $a_1, a_2, \ldots, a_{11}$

- $P(y = 0)$, $P(y = 1)$ ← while these can be learned from the dataset, they are usually just set to 0.5 each.

- $P(a_1 = 0)$, $P(a_1 = 1)$, ...
  $P(a_{11} = 0)$, $P(a_{11} = 1)$ ← It turns out that we don't need to learn these at all.

Putting it all together

New Document

Convert to bag of words

tokenize → Split into words

clean → remove punctuation, capitalization, ...

Vectorize using existing dictionary

→ Calculate (P1) $P(y=0 \mid a_1=0, \ldots a_{11}=1)$

→ Calculate (P2) $P(y=1 \mid a_1=0, \ldots, a_{11}=1)$

Prediction = Max (P_1, P_2)

Summary

- The difference between probability and statistical inference.

$$P(H|E) \neq P(E|H)$$

- When the probability you want is hard to calculate, try Bayesian inversion!
- Text can be represented by lists of numbers - the standard machine learning techniques apply.
- Naive Bayes is a surprisingly good classifier.

but is it machine learning?