

CMPSCI 240: Reasoning about Uncertainty

Lecture 16: Spam Filtering and Naive Bayes Classification

Andrew McGregor

University of Massachusetts

Outline

- 1 Review
- 2 Spam Filtering
- 3 Examples of NBC
- 4 Problems with NBC

Review

- **Total Probability** If A_1, \dots, A_n partition Ω then for any event B :

$$P(B) = P(B|A_1)P(A_1) + \dots + P(B|A_n)P(A_n) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

- **Bayes Theorem** If A_1, \dots, A_n partition Ω then for any event B :

$$P(A_i|B) = \frac{P(A_i \cap B)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)}$$

Outline

- 1 Review
- 2 Spam Filtering**
- 3 Examples of NBC
- 4 Problems with NBC

Spam Filtering

- When you receive an email you have two “hypotheses”:

$$H_1 = \text{email is spam} \quad H_2 = \text{email is not spam}$$

and note that these are partitioning events.

- You have some “observed data” e.g.,

$$D = \text{email contains the word 'cash'}$$

- How do you use the observed data to pick one of the hypotheses?
- We want to pick the “maximum a posteriori hypothesis” (MAP), i.e., the hypothesis H_i that maximizes $P(H_i|D)$.

Using Bayes Theorem for Spam Filtering

- By Bayes theorem:

$$P(H_i|D) = \frac{P(H_i \cap D)}{P(D)} = \frac{P(D|H_i)P(H_i)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)}$$

- We need to know the **prior** probabilities of H_1 and H_2 :

$$P(H_1) = ? \quad P(H_2) = ?$$

i.e., before we knew anything about the email, how likely was the email to be spam or not.

- We also need to know that **likelihood** of the observed data given each hypothesis:

$$P(D|H_1) = ? \quad P(D|H_2) = ?$$

i.e., for each hypothesis, what's the probability that we would see the observed data?

Computing Priors and Likelihoods

- To compute priors and likelihoods we use training or historical data.
- Suppose that we have analyzed 1000 pervious emails and found:

700 of the emails were spam

300 of the emails were not spam

From this, it would be natural to believe that 70% of future emails were spam and that 30% of future emails were not spam. Hence, we set the priors to be $P(H_1) = 0.7$ and $P(H_2) = 0.3$.

- Suppose we have also found that

350 of the spam emails include the word “cash”

100 of the non-spam emails include the word “cash”

From this, it's natural to set the likelihoods to be $P(D|H_1) = 1/2$ and $P(D|H_2) = 1/3$.

Putting it all together

- If $P(H_1) = 0.7$, $P(H_2) = 0.3$, $P(D|H_1) = 1/2$, $P(D|H_2) = 1/3$ then:

$$\begin{aligned} P(H_1|D) &= \frac{P(H_1 \cap D)}{P(D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} \\ &= \frac{1/2 \times 0.7}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} \\ &= \frac{0.35}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} \end{aligned}$$

and similarly

$$\begin{aligned} P(H_2|D) &= \frac{0.3 \times 1/3}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} \\ &= \frac{0.1}{P(D|H_1)P(H_1) + P(D|H_2)P(H_2)} \end{aligned}$$

- Therefore the MAP hypothesis is H_1 .

Combining Observed Data

- Suppose we have two pieces of observed data:

D_1 = email contains the word 'cash'

D_2 = email contains the word 'pharmacy'

- The MAP hypothesis is the one maximizing $P(H_j|D_1 \cap D_2)$ where

$$P(H_j|D_1 \cap D_2) = \frac{P(D_1 \cap D_2|H_j)P(H_j)}{P(D_1 \cap D_2|H_1)P(H_1) + P(D_1 \cap D_2|H_2)P(H_2)}$$

Naive Bayes Classification

- In Naive Bayes Classification (NBC) we assume that the observed data is independent conditioned on either of the hypotheses, i.e.,

$$P(D_1 \cap D_2 | H_1) = P(D_1 | H_1)P(D_2 | H_1)$$

$$P(D_1 \cap D_2 | H_2) = P(D_1 | H_2)P(D_2 | H_2)$$

- And therefore NBC picks the hypothesis H_j that maximizes

$$P(D_1 \cap D_2 | H_j)P(H_j) = P(D_1 | H_j)P(D_2 | H_j)P(H_j)$$

- Can compute the priors $P(H_1), P(H_2)$ and the likelihoods $P(D_1 | H_1), P(D_2 | H_1), P(D_1 | H_2), P(D_2 | H_2)$ from the training data.

Picking an Hypothesis

- Suppose we have multiple hypotheses H_1, H_2, \dots, H_k and we have **priors** for these hypotheses $P(H_1), P(H_2), \dots, P(H_k)$.
- If we observe some event D , the **MAP hypothesis** is the hypothesis H_i that maximizes

$$P(H_i|D) = \frac{P(D|H_i)P(H_i)}{\sum_j P(D|H_j)P(H_j)}$$

To find this we need **likelihoods** $P(D|H_j)$ for each hypothesis H_j .

- If we observe multiple events D_1, D_2, \dots, D_t , the **MAP hypothesis** is the hypothesis H_i that maximizes

$$P(H_i|D_1 \cap D_2 \cap \dots \cap D_t) = \frac{P(D_1 \cap D_2 \cap \dots \cap D_t|H_i)P(H_i)}{\sum_j P(D_1 \cap D_2 \cap \dots \cap D_t|H_j)P(H_j)}$$

- In **Naive Bayes Classification** we assume that for each j :

$$P(D_1 \cap D_2 \cap \dots \cap D_t|H_j) = P(D_1|H_j)P(D_2|H_j) \dots P(D_t|H_j)$$

Outline

- 1 Review
- 2 Spam Filtering
- 3 Examples of NBC**
- 4 Problems with NBC

Understanding Happiness

Quora

Ask or Search Quora

Ask Question

Read

Answer

Notifications

9

Barna

IBM Thomas J. Watson Research Center

IBM Research

+9



How can a computer understand which story is happy and which one is sad?

Or it doesn't, what could be done and how?

Request 5 Follow 5 Comment Share 1 Downvote

...

Have this question too? Request Answers:



Request From Quora

We will distribute this question to writers, and notify you about new answers.



Dan Zhang, Computer Engineering PhD student at UT Austin; interned at Microsoft ...
169 Answers in Computer Programming



Luis Argerich, Data Science Professor at the University of Buenos Aires (UBA)
233 Answers in Machine Learning



Anima Anandkumar, Visiting researcher
16 Answers in Machine Learning



View More or Search

Related Questions

How can we convert a one line sad story to a two liner happy one?

How can one make someone that is sad, happy?

Can sadness be happiness?

What do I do when I can't understand why I'm feeling happy/sad/depressed?

What works better in a sad mood, inspiring and happy stories or sadder stories of other people?

Why is it so much easier to make up stories about sad, depressing things rather than happy?

Can happiness exist without sadness?

Why all French songs are sad? (I can't find any happy one)?

How does a good understanding of computer science help one to understand computer programming?

Why do I like sad/love songs more than happy ones?

Understanding Happiness

- There are 1000 documents out of which 700 are sad documents and 300 are happy.
 - Out of the happy documents, 100 contain the word *happy*.
 - Out of the sad documents, 100 contain the word *happy*.
 - Out of the happy documents, 50 contain the word *shock*.
 - Out of the sad documents, 350 contain the word *shock*.
- Computer is given a document which has both the words *happy* and *shock*. Find the maximum a posteriori hypothesis.

Combing Evidence: Bird Watching

- You see a blue parrot and you hypothesize that it's a Norwegian Blue
- Your birdwatching book says:
 - Only 10% of blue parrots are Norwegian Blues
 - Norwegian Blues spend 60% of their time lying down
 - Other blue parrots only spend 20% of their time lying down
 - 80% of Norwegian Blues have lovely plumage
 - 20% of other blue parrots have lovely plumage.
- If the parrot is lying (D_1) and has lovely plumage (D_2): what is the probability that it's a Norwegian Blue (N)?

$$P(N|D_1 \cap D_2) = \frac{P(D_1 \cap D_2|N) P(N)}{P(D_1 \cap D_2|N) P(N) + P(D_1 \cap D_2|N^c) P(N^c)}$$

- Not enough information for $P(D_1 \cap D_2|N)$ & $P(D_1 \cap D_2|N^c)$
- Could use NBC: Assume $P(D_1 \cap D_2|N) = P(D_1|N) P(D_2|N)$ and $P(D_1 \cap D_2|N^c) = P(D_1|N^c) P(D_2|N^c)$

Outline

- 1 Review
- 2 Spam Filtering
- 3 Examples of NBC
- 4 Problems with NBC**

Problems with NBC: Lack of Independence

- Features might not be conditionally independent
- Suppose we are doing spam filtering and

$$D_1 = \{\text{email includes the word 'western'}\}$$

$$D_2 = \{\text{email includes the word 'union'}\}$$

$$S = \{\text{email is spam}\}$$

- Say 10% of email is spam

Spam: 20% includes “western”, “union”, “western union”

Non-Spam: 5% includes “western”, 5% “union”, 4% “western union”

Problems with NBC: Lack of Independence Continued

- Suppose you see an email that includes “western union”

$$P(S|D_1 \cap D_2) = \frac{P(D_1 \cap D_2|S) P(S)}{P(D)} = \frac{0.2 \times 0.1}{P(D)} = \frac{0.02}{P(D)}$$

$$P(S^c|D_1 \cap D_2) = \frac{P(D_1 \cap D_2|S^c) P(S^c)}{P(D)} = \frac{0.04 \times 0.9}{P(D)} = \frac{0.036}{P(D)}$$

so we would conclude the email isn't spam.

- NBC would do the following:

$$P(S|D_1 \cap D_2) \approx \frac{P(D_1|S) P(D_2|S) P(S)}{P(D)} = \frac{0.2 \times 0.2 \times 0.1}{P(D)} = \frac{0.004}{P(D)}$$

$$P(S^c|D_1 \cap D_2) \approx \frac{P(D_1|S^c) P(D_2|S^c) P(S^c)}{P(D)} = \frac{0.05 \times 0.05 \times 0.9}{P(D)} = \frac{0.00225}{P(D)}$$

so we would conclude the email is spam.

The Effect of Assuming Independence

- Suppose you have two hypotheses:

$$H_1 = \text{"aliens have invaded Amherst"}$$

$$H_2 = \text{"aliens have not invaded Amherst"}$$

with priors $P(H_1) = 1/100$ and $P(H_2) = 99/100$.

- Let D_i be the event that your i th friend tells you that aliens have landed. Suppose $P(D_i|H_1) = 1$ and $P(D_i|H_2) = 0.1$
- Are you more likely to conclude aliens have landed if you believe your friends are independent conditioned on your hypothesis?

Problems with NBC: Using Rare or Too Common Features

- Suppose we pick a very common word as one of the features
- If it always occurs in a spam message (in your training set) but not always in your non-spam messages, you'll always conclude an email without that word is non-spam.
- Possible solutions:
 - 1 Train, train, train! Use as much training data as possible.
 - 2 Ignore words that don't both occur and fail to occur in some positive example and some negative example.
 - 3 "Smooth" the data: if "hippo" occurred in 17 of the positive examples and none of the negative examples, pretend the counts are 18 and 1.