

Naive Bayes for Sentiment Analysis

Yannet Interian

Agenda

- Review of Naive Bayes and Sentiment Analysis
- Coding Naive Bayes in Spark

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.

Important commercial application



Naive Bayes Intuition

- Simple (“naive”) classification method based on Bayes rule
- Relies on very simple representation of document
 - Bag of words / unigram language model

The bag of words representation

$$F(\text{I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.}) = C$$


The bag of words representation

$$F(\text{I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet.}) = C$$


The bag of words representation: using a subset of words

$$F(\text{document}) = C$$

The document text is:

```
x love xxxxxxxxxxxxxxxxxxxx sweet  
xxxxxxxx satirical xxxxxxxxxxxx  
xxxxxxxxxxxx great xxxxxxxx  
xxxxxxxxxxxxxxxxxxxxxxxxxxxx fun xxxx  
xxxxxxxxxxxxxxxxxxxx whimsical xxxx  
romantic xxxx laughing  
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx  
xxxxxxxxxxxxxxxxxxxx recommend xxxxxx  
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx  
xx several xxxxxxxxxxxxxxxxxxxxxxxx  
xxxxxx happy xxxxxxxxxxxx again  
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx  
xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
```

The result C is represented by two smiley faces:



- 😊 (Yellow happy face)
- 😞 (Green sad face)

The bag of words representation

F (

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

) = **C**

Bayes' Rule Applied to Documents and Classes

For a document d and a class c

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

Naive Bayes Classifier (I)

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

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

Dropping the denominator

likelihood

prior

Naïve Bayes Classifier (II)

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

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

Document d
represented as
features
 $x_1 \dots x_n$

Multinomial Naïve Bayes

Independence Assumptions

- Bag of Words assumption: Assume position doesn't matter
- Example for $x_i = \{ \text{occurrence of word } \textit{like} \}$
- Conditional Independence: Assume the features $x_i | c$ are independent for every class c .

$$P(x_1, \dots, x_n | c) = P(x_1 | c) \cdot P(x_2 | c) \cdot P(x_3 | c) \cdot \dots \cdot P(x_n | c)$$

Multinomial Naïve Bayes Classifier

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

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

$P(x|c)$ "how much
evidence x contributes
that c the correct class"

Naïve Bayes Learning

Learning the Multinomial Naïve Bayes Model

First attempt: simply use the frequencies in the data

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

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

What is the problem with this attempt?

Laplace (add-1) smoothing for Multinomial Naïve Bayes

- What if we have seen no training documents with the word "fantastic" and classified in the topic positive?

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

- Add-1 Smoothing

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

Multinomial Naïve Bayes Learning: summary

For every class

- compute priors

N = Number of documents

N_c = Number of documents in class c

$$\hat{P}(c) = \frac{N_c}{N}$$

For every word w and
every class c

V = Set of unique words

$count(w, c)$ = Frequency of w in c

$count(c)$ = Number of words in c

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

Find c such that:

$$\operatorname{argmax} \hat{P}(c) \hat{P}(d|c)$$

What is the time complexity of this algorithm?

Naïve Bayes: unknown words

- If your training set is expected to have unknown words: add a new word w_u to the vocabulary.

$$\begin{aligned}\hat{P}(w_u | c) &= \frac{\text{count}(w_u, c) + 1}{\left(\sum_{w \in V} \text{count}(w, c) \right) + |V + 1|} \\ &= \frac{1}{\left(\sum_{w \in V} \text{count}(w, c) \right) + |V + 1|}\end{aligned}$$

Summary

`count(w, pos)`

`count(w, neg)`

`count(pos)` **and** `count(neg)`

`V` number of unique words in the training set

$$P(w | c) = (\text{count}(w, c) + 1) / (\text{count}(c) + V + 1)$$

Given that how do we compute the class of a document?

$$P(c | d)$$

Summary

`count(w, pos)`

`count(w, neg)`

`count(pos)` **and** `count(neg)`

`V` number of unique words in the training set

$$P(w|c) = [\text{count}(w, c) + 1] / (\text{count}(c) + V + 1)$$

Given that how do we compute the class of a document?

$$\mathbf{log} P(c|d) \sim \text{sum_i} \mathbf{log} P(w_i|c) - P(c)$$

Acknowledgment

Some of these slides are adapted from the NLP class from coursera.org taught by the Stanford professors: [Dan Jurafsky](#) and [Chris Manning](#).
<https://class.coursera.org/nlp/>