Text Classification - I

Pawan Goyal

CSE, IIT Kharagpur

Week 11, Lecture 4

Example: 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

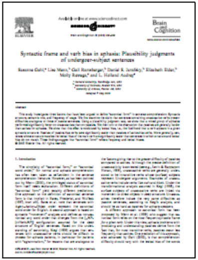
- It was pathetic. The worst part about it was the boxing scenes.

Example: Male or Female Author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam...
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets...

Example: 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
- ...

Taxt Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language identification
- Sentiment analysis
- ...

Text classification: problem definition

Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$

Text classification: problem definition

Input

- A document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$

Output

A predicted class $c \in C$

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

black-list-address OR ("dollars" AND "have been selected")

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features

Spam

black-list-address OR (“dollars” AND “have been selected”)

Pros and Cons

Accuracy can be high if rules carefully refined by expert, *but building and maintaining these rules is expensive.*

Classification Methods: Supervised Machine Learning

- Naïve Bayes
- Logistic regression
- Support-vector machines
- ...

Naïve Bayes Intuition

- Simple classification method based on Bayes' rule
- Relies on very simple representation of document - Bag of words

Bag of words for document classification

Test document

parser
language
label
translation
...

?

Machine Learning	NLP	Garbage Collection	Planning	GUI
learning	parser	garbage	planning	...
training	tag	collection	temporal	
algorithm	training	memory	reasoning	
shrinkage	translation	optimization	plan	
network...	language...	region...	language...	

Bayes' rule for documents and classes

For a document d and a class c

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

Bayes' rule for documents and classes

For a document d and a class c

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

Naïve Bayes Classifier

$$\begin{aligned} c_{MAP} &= \arg\max_{c \in C} P(c|d) \\ &= \arg\max_{c \in C} P(d|c)P(c) \\ &= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n|c)P(c) \end{aligned}$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i | c_j)$ are independent given the class c_j.

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

Naïve Bayes classification assumptions

$$P(x_1, x_2, \ldots, x_n | c)$$

Bag of words assumption

Assume that the position of a word in the document doesn't matter

Conditional Independence

Assume the feature probabilities $P(x_i|c_j)$ are independent given the class c_j.

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

$$c_{NB} = \arg\max_{c \in C} P(c) \prod_{x \in X} P(x|c)$$

Learning the model parameters

Maximum Likelihood Estimate

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

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

Learning the model parameters

Maximum Likelihood Estimate

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

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

Problem with MLE

Suppose in the training data, we haven't seen the word "fantastic", classified in the topic 'positive'.

$$\hat{P}(fantastic|positive) = 0$$

Learning the model parameters

Maximum Likelihood Estimate

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

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

Problem with MLE

Suppose in the training data, we haven't seen the word "fantastic", classified in the topic 'positive'.

$$\hat{P}(fantastic|positive) = 0$$

$$c_{NB} = \arg\max_c \hat{P}(c) \prod_{x\in X} \hat{P}(x_i|c)$$

Laplace (add-1) smoothing

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

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