Computational Social Science Methods in R
Text Classification

Taehee Kim

Summer Semester 2022

University of Oldenburg

Basics of document analysis[1]

Tokenization

- Segmenting an input stream into an ordered sequence of tokens is called tokenization.
- A system which splits texts into word tokens is called a tokenizer.
- Example:

  Input text: "Tony likes pizza and Mike likes pasta."

  output: {"Tony", "likes", "pizza", "and", "Mike", "likes", "pasta", "."}

Text Normalization: Stemming and Lemmatization

- different forms of a word: e.g., organize, organizes, and organizing
- The goal of both stemming and lemmatization is to reduce inflectional and derivationally related forms of a word to a common base form
- Example: 'the boy's cars are different colors' → the boy car be differ color

[1] Tokenization: https://nlp.stanford.edu/IR-book/html/htmledition/tokenization-1.html, Text Normalization: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Stemming

- Stemming: usually refers to chop off the ends of words (often includes the removal of derivational affixes)
- beautifully $\rightarrow$ beauti

Lemmatization

- Lemmatization: usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the lemma.
- beautifully $\rightarrow$ beautifully

Stop words

- refers common words in a language so that does not contain important significance to be used in Search Queries.
- in English: as, the, is, are, with etc..

Tokenization and Stemming

German

- Python NLTK package → install
- punkt in modules install

Text classification

Goal

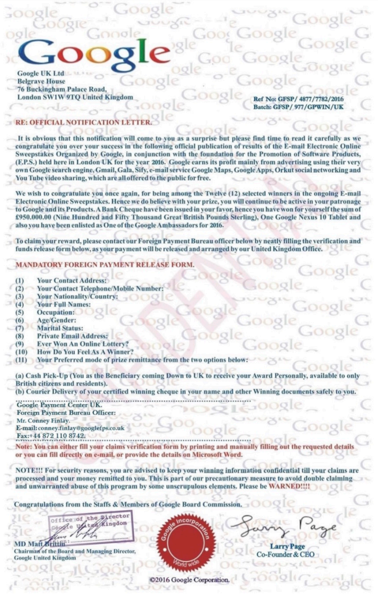
- Assign 'texts' from a universe to two or more classes or categories

Applications

- Spam vs. Non-spam email
- Email filtering
- News events tracked and filtered by topics
- Journal articles indexed by subject categories

Who wrote which Federalist papers?

- 1987: anonymous essays try to convince New York to ratify U.S. Constitution
- Jay, Madison, and Hamilton
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods


Google UK Ltd
Belgrave House
76 Buckingham Palace Road,
London SW1W 9TQ United Kingdom

Ref No: GFSP/ 4877/7782/2016
Batch: GFSP/ 977/GPWIN/UK

RE: OFFICIAL NOTIFICATION LETTER.

It is obvious that this notification will come to you as a surprise but please find time to read it carefully as we congratulate you over your success in the following official publication of results of the E-mail Electronic Online Sweepstakes Organized by Google, in conjunction with the foundation for the Promotion of Software Products, (F.P.S.) held here in London UK for the year 2016. Google earns its profit mainly from advertising using their very own Google search engine, Gmail, Gala, Sify, e-mail service Google Maps, Google Apps, Orkut social networking and You Tube video sharing, which are all offered to the public for free.

We wish to congratulate you once again, for being among the Twelve (12) selected winners in the ongoing E-mail Electronic Online Sweepstakes. Hence we do believe with your prize, you will continue to be active in your patronage to Google and its Products. A Bank Cheque have been issued in your favor, hence you have won for yourself the sum of £950.000.00 (Nine Hundred and Fifty Thousand Great British Pounds Sterling), One Google Nexus 10 Tablet and also you have been enlisted as One of the Google Ambassadors for 2016.

To claim your reward, please contact our Foreign Payment Bureau officer below by neatly filling the verification and funds release form below, as your payment will be released and arranged by our United Kingdom Office.

MANDATORY FOREIGN PAYMENT RELEASE FORM.

(1) Your Contact Address:
(2) Your Contact Telephone/Mobile Number:
(3) Your Nationality/Country:
(4) Your Full Names:
(5) Occupation:
(6) Age/Gender:
(7) Marital Status:
(8) Private Email Address:
(9) Ever Won An Online Lottery?
(10) How Do You Feel As A Winner?
(11) Your Preferred mode of prize remittance from the two options below:

(a) Cash Pick-Up (You as the Beneficiary coming Down to UK to receive your Award Personally, available to only British citizens and residents).
(b) Courier Delivery of your certified winning cheque in your name and other Winning documents safely to you.

Google Payment Center UK.
Foreign Payment Bureau Officer:
Mr. Conney Finlay.
E-mail: conney.finlay@googlefps.co.uk
Fax:+44 872 110 8742.

Note: You can either fill your claims verification form by printing and manually filling out the requested details or you can fill directly on e-mail, or provide the details on Microsoft Word.

NOTE!!! For security reasons, you are advised to keep your winning information confidential till your claims are processed and your money remitted to you. This is part of our precautionary measure to avoid double claiming and unwarranted abuse of this program by some unscrupulous elements. Please be WARNED!!!!

Congratulations from the Staffs & Members of Google Board Commission.

Office of the Director
Google United Kingdom
Sign:

MD Matt Brittin
Chairman of the Board and Managing Director,
Google United Kingdom

Google Incorporation
World wide

Larry Page
Co-Founder & CEO

©2016 Google Corporation.

Figure 1: Example of a spam email

Hand-coded rules

- Rules based on combinations of words or other features
  - Spam mail: black-list-address OR (“winning" AND “payment")
- Accuracy can be high if rules are carefully refined by experts
- However, building and maintaining these rules are expensive

Supervised machine learning

Input

- Document D
- A fixed set of classes C
- A training set (hand-labeled documents)

Output

- A learned classifier

Naive Bayes

- One of the most important classifiers
- Relies on bag of words

Bag of words approach

- does not consider the order of words
- but the counts of individual words
- works for multiple classification

Example

I will vote for XXX party! I agree with their ideas and support their policies. Not only they are sophisticated and smart but also sincere PATRIOT!

I will vote for XXX party! I agree with their ideas and support their policies. Not only they are sophisticated and smart but also sincere PATRIOT!

Naive Bayes[2]

Bayes Rule

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Text Classification Naive Bayes (Multinomial)

$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$, where d: a document and c: a class

We compute $P(c_i|d)$, where i =1,...,n (number of class)

and adopt c_i which maximize $P(c_i|d)$

[2]Recommended reading: Raschka, Sebastian. "Naive Bayes and Text Classification I– Introduction and Theory." arXiv preprint arXiv:1410.5329 (2014)

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

$P(d|c_i)P(c_i)$

$= P(w_1, w_2, ..., w_n|c_i)P(c_i)$

$= P(w_1|c_i)P(w_2|c_i)...P(w_n|c_i)P(c_i)$, where w: word, n: number of words in d

- $P(c_i) = \frac{(d,c_i)}{\sum(d,c)} = \frac{\text{number of doc. in class i}}{\text{number of all doc.}}$

- $P(w_j|c_i) = \frac{(w_j,c_i)}{\sum(w,c_i)} = \frac{\text{number of word j in class i}}{\text{number of all words in class i}}$

Laplace Smoothing

$P(w_j|c_i) = \frac{(w_j,c_i)+1}{\sum(w,c_i)+|V|}$

V : number of all unique words appeared in training documents

Example

Training:

- dog, cat, cat $\rightarrow$ class A
- dog, dog, bird $\rightarrow$ class A
- dog, fox $\rightarrow$ class A
- fish, dolphin, dog $\rightarrow$ class B

Test

- dog, dog, dog, fish, dolphin $\rightarrow$???

Support Vector Machines (SVM)

- SVM is one of the popular machine learning algorithms since it shows good performance in a diverse types of data, including text
- it determines the best decision boundary between vectors that belong to a given class
- SVM shows good performance with data that have lots of features $\rightarrow$ such as text data
- most text classification problems are linearly separable, and linear Kernel shows good performance for data with a lot of features, such as text data

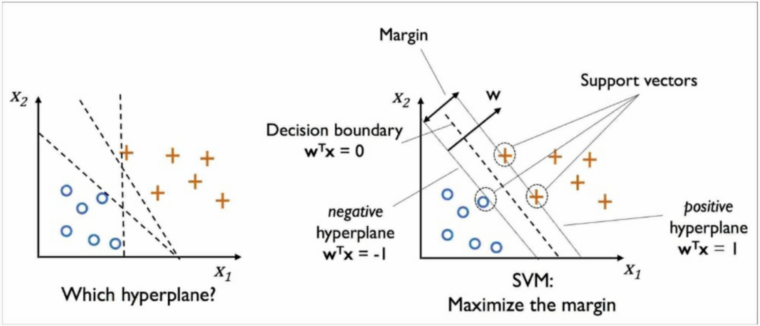


Figure 2: Support Vector Machine

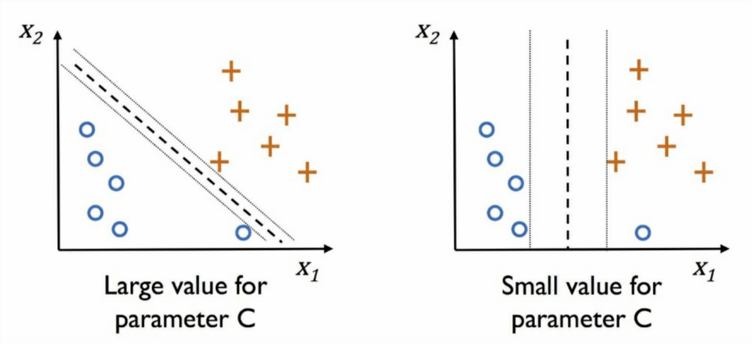


Figure 2: Support Vector Machine

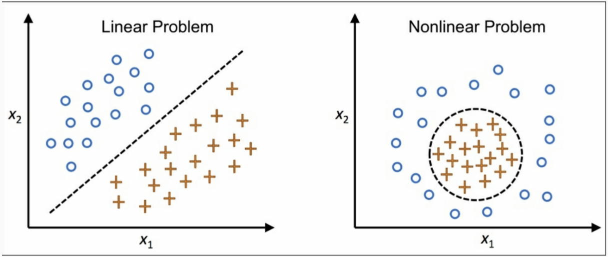


Figure 3: Linear and Nonlinear Problem

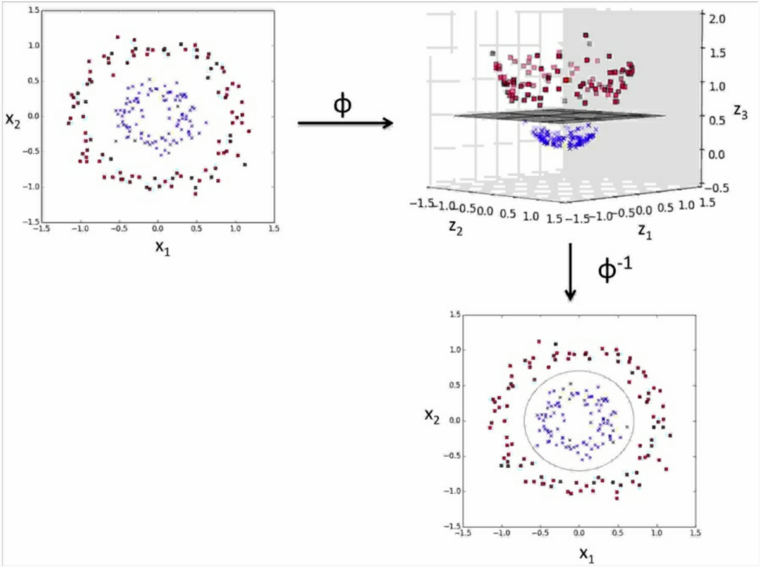


Figure 3: Linear and Nonlinear Problem

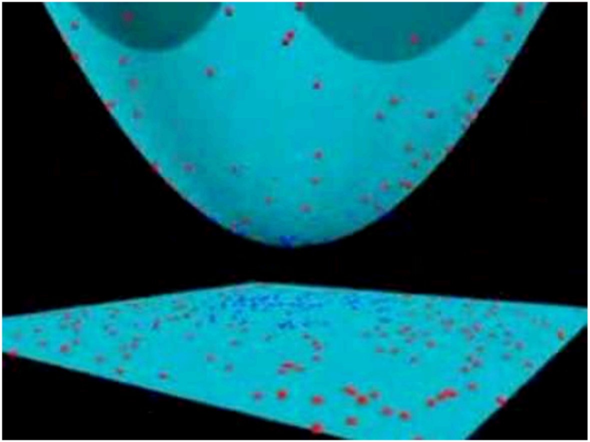


Figure 4: https://www.youtube.com/watch?v=3liCbRZPrZA

Text vectorization

- transform text into a vector
- define a fixed length vector where each entry corresponds to a word appeared in training texts.
- vector entries: count a word appeared in the text

Example

- Suppose that we have following words in training texts

  (I, do, not, am, like, happy, pizza)
- We vectorize the text “I like pizza"

  (1, 0, 0, 0, 1, 0, 1)

Term frequency–inverse document frequency(tf–idf)

- numerical statistic which represent the importance of a word
- it is often used as a weight

Term frequency

- The term frequency $tf_{t,d}$ of term t in document d is defined as the number of times that t occurs in d
- A document with 10 occurrences of the term is more relevant than a document with one occurrence of the term (but the relevance does not increase proportionally).

Document frequency

- Are rare terms or frequent terms more informative? $\rightarrow$ rare terms
- we want to a high weight for rare terms
- df_t is the document frequency, the number of documents that t occurs in, N is the total number of documents

Inverse document frequency

- a measure of the informativeness of the term.
- $idf_t = log_{10} \frac{N}{df_t}$

TF–IDF

- We assign a tf-idf weight for each term t in each document d:
- $w_{t,d} = (1 + log\ tf_{t,d}) \cdot log \frac{N}{df_t}$

  N: total number of documents

  Increases with the number of occurrences within a document

  Increases with the rarity of the term in the collection