Indian Institute of Technology Kharagpur Department of Computer Science & Engineering

CS60075 Natural Language Processing Autumn 2020

Module 4: Part 1
Classification tasks in NLP

Sep 23 2020

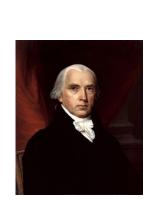
The text classification problem

- Given a text $d = (w_1, w_2, ..., w_T) \in V^*$
- choose a label $c_i \in C$

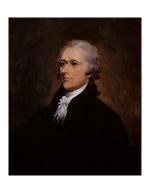
- Some applications
 - Sentiment Analysis {pos, neg, neutral}
 - Spam filtering
 - Language identification
 - Age/gender identification
 - Authorship identification

Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods



James Madison



Alexander Hamilton

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...
- 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...

Positive or negative movie review?







 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.



What is the subject of this article?

MEDLINE Article







Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

Susanna Gahl," Lies Mann," Gail Rameberger," Daniel S. Jurafeky," Elizabeth Elder," Molly Revega," and L. Holland Audrey"

Americanus, factogo as, atr samely of factors, make the atr samely of classe, from all, after

This well, interligent from form for loss that expert is also benefited from its measure complication for each source, according to the self-ligence in each of the first of the self-ligence in the self-ligence in the self-ligence according to the self-ligence in the self-ligence according to the self-lige the series in orders. We stop the tits often is makinged to being like, i.e. the facilities for a cert arranged to stop spreads arrange. Nation of positioning series are significantly under the position of activities webs. Were greening serthe analysis are also regarded to be and the set of the set of the set of the set of the second to be a second His driver result. These hadrops aggree that "extended form" refers beginning and halled blass. © 2013 Theories A. All rights married.

The display of frequency larger or frequency word order," for named and aphasis comprehension has often been taken as self-existed in the sentence comprehension librarium. However, as has been policied out by Menn (2005), the printinged status of experied form half reads explanation. Different definitions of "severine form" yield tending different predictions. One approach to the definition of severine sentence Seen in that Implicit in Room, Proceeds, and Wolfes-(1885, how all). Being et al., note that genterous with Agent, Anther, Digital under represent the securiorists and order for English, A second agreement is best for greate "moment" and so of differ a remain-tended any used order the diarge from the LATA. (offerbAF) and position around for the day arrange of English sentence. Read on the units. standing of secondary, Keyl (1995) expensivel area prome for agreein parteria, in particular for parteria

See that, for most transitive ratio, positive owner has frequently than certain. One production of this approach, also advanced by Calif (2012), is that comprehension

COLUMN TO THE WAY THE PARTY OF THE PARTY OF THE PARTY.

alon feeder halos to my over difficulty to should be as hard as possible for agharic speciers. A different approach to paramise from his been proposal by Mann et al. (1998) who suggest that me saried from relies on the most frequent explosive frame producing and understanding popular darket from the

differing should vary with the below bloom? the words

the factors giving the justice product illigating of positions property to palvay. Although the product definition of

community is contained puring, Levin & Represent Rome, 1985, community with an approach units.

stand to be impositive only whose (softest subjects represent Undergott organization Exemples of seasons).

salter onto include onto the man and state. Under the transformational analysis assumed in Keyl (1998), the

serior calibra of manuality toda on Irisal da

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology



Classification Methods: Supervised Machine Learning

• Input:

- a document d
- a fixed set of classes $C = \{c_1, c_2, ..., c_J\}$
- A training set of m hand-labeled documents $(d_1, c_1), \dots, (d_m, c_m)$

• Output:

• a learned classifier $\gamma: d \rightarrow c$

The bag-of-words

- One challenge: the sequential representation (w1, w2, . . . , wT) may have a different length T for every document.
- The bag-of-words is a fixed-length representation, which consists of a vector of word counts:

- w = It was the best of times, it was the worst of times
- x =[aardvark, . . . , best, . . . , it , . . . , of , . . . , zyther]

• The length of x is equal to the size of the vocabulary, V.

Bayes' Rule Applied to Documents and Classes

For a document d and a class C

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

Naïve Bayes Classifier (I)

$$c_{MAP} = \operatorname*{argmax} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

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

Bayes Rule

$$= \operatorname*{argmax} P(d \mid c) P(c)$$

$$c \in C$$

Dropping the denominator

Naïve Bayes Classifier (II)

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

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

Document d represented as features x1..xn

But where will we get these probabilitites?

Multinomial Naïve Bayes Independence Assumptions

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

- Bag of Words assumption
 - Assume position doesn't matter
- Conditional Independence
 - Assume the feature probabilities $P(x_i|c)$ are independent given the class c.

$$P(x_1,...,x_n | c) = P(x_1 | c) \bullet P(x_2 | c) \bullet P(x_3 | c) \bullet ... \bullet P(x_n | c)$$

Multinomial Naïve Bayes Classifier

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

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

This is why it's naïve!

[Jurafsky and Martin]

Learning the Multinomial Naïve Bayes Model

- Maximum likelihood estimates
 - 1. simply use the frequencies in the data

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

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

Parameter Estimation

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

fraction of times word w_i appears among all words in documents of topic c_j

- Create mega-document for topic j by concatenating all docs in this topic
 - Use frequency of w in mega-document

Problem with Maximum Likelihood

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

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

 Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \operatorname{argmax}_{c} \hat{P}(c) \prod_{i} \hat{P}(x_{i} \mid c)$$

Laplace Smoothing

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

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

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

Multinomial Naïve Bayes: Learning

From training corpus, extract Vocabulary

- Calculate $P(c_i)$ terms
 - For each c_i in C do $docs_i \leftarrow all docs with class = c_i$

$$P(c_j) \leftarrow \frac{|docs_j|}{|total \# documents|}$$

Calculate $P(w_k \mid c_i)$ terms

- $Text_i \leftarrow single doc containing all <math>docs_i$
- For each word w_k in *Vocabulary* $n_k \leftarrow \#$ of occurrences of w_k in $Text_i$

$$P(c_{j}) \leftarrow \frac{|docs_{j}|}{|\operatorname{total} \# \operatorname{documents}|} \qquad P(w_{k} | c_{j}) \leftarrow \frac{n_{k} + \alpha}{n + \alpha |\operatorname{Vocabulary}|}$$

Naive Bayes

- Very fast, low storage requirements
- Robust to irrelevant features
- A good, dependable baseline for text classification
 - But other classifiers give better accuracy