

Text Classification and Naïve Bayes

The Task of Text Classification

Many slides are adapted from slides by Dan Jurafsky

Is this spam?

Subject: Important notice!

From: Stanford University <newsforum@stanford.edu>

Date: October 28, 2011 12:34:16 PM PDT

To: undisclosed-recipients;;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

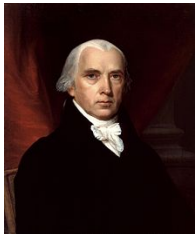
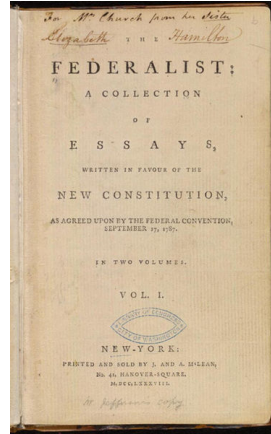
<http://www.123contactform.com/contact-form-StanfordNew1-236335.html>

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information about the new services.

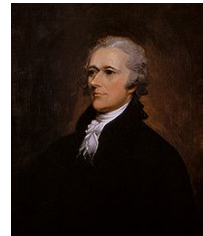
© Stanford University. All Rights Reserved.

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 Cochinchina; 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...

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

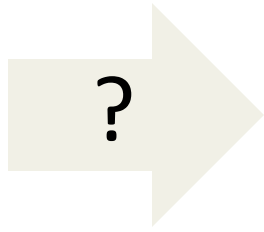


- It was pathetic. The worst part about it was the boxing scenes.

What is the subject of this article?

MEDLINE Article

The screenshot shows the top portion of a MEDLINE article. At the top left is the Elsevier logo. To its right is the text 'Available at www.sciencedirect.com' and the 'Brain & Cognition' journal logo. The article title is 'Syntactic frames and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences'. Below the title are the authors' names: Susanne Güll, Lisa Mann, Gail Ramlinger, Daniel S. Jantsky, Elisabeth Röder, Molly Rawaga, and L. Harlan Aulay. The article is published in 'Brain & Cognition', volume 100, pages 1-10, in 2015. The abstract begins with 'The study investigates how verbs that have been argued to define "lexical slots" in sentence comprehension (syntax)...



MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Text Classification

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

Text Classification: definition

- *Input*:
 - a document d
 - a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
- *Output*: a predicted class $c \in C$

Classification Methods:

Hand-coded rules

- Rules based on combinations of words or other features
 - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
 - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

- *Input:*
 - a document d
 - a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
 - A training set of m hand-labeled documents $(d_1, c_1), \dots, (d_m, c_m)$
- *Output:*
 - a learned classifier $\gamma: d \rightarrow c$

Classification Methods: Supervised Machine Learning

- Any kind of classifier
 - Naïve Bayes
 - Logistic regression, maxent
 - Support-vector machines
 - k-Nearest Neighbors
 - ...

Text Classification and Naïve Bayes

The Task of Text Classification

Text Classification and Naïve Bayes

Formalizing the Naïve Bayes Classifier

Naïve Bayes Intuition

- Simple (“naïve”) classification method based on Bayes rule
- Relies on very simple representation of document
 - Bag of words

Bayes' Rule Applied to 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 (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

Naïve Bayes Classifier (II)

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

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

Document d
represented as
features
 $x_1 \dots x_n$

Naïve Bayes Classifier (IV)

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

$O(|X|^n \cdot |C|)$ parameters

How often does this class occur?

Could only be estimated if a very, very large number of training examples was available.

We can just count the relative frequencies in a corpus

Multinomial Naïve Bayes

Independence Assumptions

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

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

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

The bag of words representation

Y (

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

Y (

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

) = C

The diagram illustrates the bag of words representation. A large letter 'Y' is followed by an opening parenthesis. To the right is a table with two columns: the first column lists words and the second column lists their counts. The words are 'great', 'love', 'recommend', 'laugh', 'happy', and an ellipsis. The counts are 2, 2, 1, 1, 1, and an ellipsis. To the right of the table is a closing parenthesis, followed by an equals sign and a large letter 'C'. Below the 'C' are two icons: a thumbs up and a thumbs down.

Bag of words for document classification

?

Test document

parser
language
label
translation
...

Machine Learning

learning
training
algorithm
shrinkage
network...

NLP

parser
tag
training
translation
language...

Garbage Collection

garbage
collection
memory
optimization
region...

Planning

planning
temporal
reasoning
plan
language...

GUI

...

Applying Multinomial Naive Bayes Classifiers to Text Classification

positions \leftarrow all word positions in test document

$$c_{NB} = \operatorname{argmax}_{c_j \in \mathcal{C}} P(c_j) \prod_{i \in \text{positions}} P(x_i | c_j)$$

Text Classification and Naïve Bayes

Formalizing the Naïve Bayes Classifier

Text Classification and Naïve Bayes

Naïve Bayes: Learning

Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
 - simply use the frequencies in the data

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

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

Parameter estimation

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

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** (*thumbs-up*)?
- Zero probabilities cannot be conditioned away, no matter the other evidence!

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

$$c_{MAP} = \operatorname{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Laplace (add-1) smoothing: unknown words

Add one extra word to the vocabulary, the “unknown word” w_u

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

Underflow Prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow.
- Since $\log(xy) = \log(x) + \log(y)$
 - Better to sum logs of probabilities instead of multiplying probabilities.
- Class with highest un-normalized log probability score is still most probable.

$$c_{NB} = \operatorname{argmax}_{c_j \in C} \log P(c_j) + \sum_{i \in \text{positions}} \log P(x_i | c_j)$$

- Model is now just max of sum of weights

Text Classification and Naïve Bayes

Naïve Bayes: Learning

Text Classification and Naïve Bayes

Multinomial Naïve Bayes: A Worked Example

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

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

Priors:

$$P(c) = \frac{3}{4}$$

$$P(j) = \frac{1}{4}$$

Conditional Probabilities:

$$P(\text{Chinese}|c) = (5+1) / (8+7) = 6/15$$

$$P(\text{Tokyo}|c) = (0+1) / (8+7) = 1/15$$

$$P(\text{Japan}|c) = (0+1) / (8+7) = 1/15$$

$$P(\text{Chinese}|j) = (1+1) / (3+7) = 2/10$$

$$P(\text{Tokyo}|j) = (1+1) / (3+7) = 2/10$$

$$P(\text{Japan}|j) = (1+1) / (3+7) = 2/10$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Choosing a class:

$$P(c|d5) \propto \frac{3}{4} * \left(\frac{6}{15}\right)^3 * \frac{1}{15} * \frac{1}{15} \approx 0.0002$$

$$P(j|d5) \propto \frac{1}{4} * \left(\frac{2}{10}\right)^3 * \frac{2}{10} * \frac{2}{10} \approx 0.00008$$

Summary: Naive Bayes is Not So Naive

- Robust to Irrelevant Features

Irrelevant Features cancel each other without affecting results

- Very good in domains with many equally important features

Decision Trees suffer from *fragmentation* in such cases – especially if little data

- Optimal if the independence assumptions hold: If assumed independence is correct, then it is the Bayes Optimal Classifier for problem

- A good dependable baseline for text classification

– **But we will see other classifiers that give better accuracy**

Text Classification and Naïve Bayes

Multinomial Naïve Bayes: A Worked Example

Text Classification and Naïve Bayes

Text Classification: Evaluation

The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

Precision and recall

- **Precision:** % of selected items that are correct
Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

A combined measure: F

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- People usually use balanced F1 measure

– i.e., with $\beta = 1$ (that is, $\alpha = \frac{1}{2}$):

+R)

$$F = 2PR/(P$$

Confusion matrix c

- For each pair of classes $\langle c_1, c_2 \rangle$ how many documents from c_1 were incorrectly assigned to c_2 ?
 - $c_{3,2}$: 90 wheat documents incorrectly assigned to poultry

Docs in test set	Assigned UK	Assigned poultry	Assigned wheat	Assigned coffee	Assigned interest	Assigned trade
True UK	95	1	13	0	1	0
True poultry	0	1	0	0	0	0
True wheat	10	90	0	1	0	0
True coffee	0	0	0	34	3	7
True interest	-	1	2	13	26	5
True trade	0	0	2	14	5	10

Per class evaluation measures

Recall:

Fraction of docs in class i classified correctly:

$$\frac{c_{ii}}{\sum_j c_{ij}}$$

Precision:

Fraction of docs assigned class i that are actually about class i :

$$\frac{c_{ii}}{\sum_j c_{ji}}$$

Accuracy: (1 - error rate)

Fraction of docs classified correctly:

$$\frac{\sum_i c_{ii}}{\sum_j \sum_i c_{ij}}$$

Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- **Macroaveraging:** Compute performance for each class, then average. **Average on classes**
- **Microaveraging:** Collect decisions for each instance from all classes, compute contingency table, evaluate. **Average on instances**

Micro- vs. Macro-Averaging: Example

Class 1

	Truth: yes	Truth: no
Classifier: yes	10	10
Classifier: no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision: $(0.5 + 0.9)/2 = 0.7$
- Microaveraged precision: $100/120 = .83$
- Microaveraged score is dominated by score on common classes

Development Test Sets and Cross-validation

Training set

Development Test Set

Test Set

- Metric: P/R/F1 or Accuracy
- Unseen test set
 - avoid overfitting ('tuning to the test set')
 - more conservative estimate of performance
- Cross-validation over multiple splits
 - Handle sampling errors from different datasets
 - Pool results over each split
 - Compute pooled dev set performance

Training Set

Dev Test

Training Set

Dev Test

Dev Test

Training Set

Test Set

Text Classification and Naïve Bayes

Text Classification: Evaluation