NLP: Text Classification I

Katja Markert

School of Computing
University of Leeds
markert@comp.leeds.ac.uk

October 23, 2013

Today

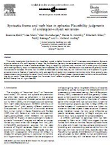
1. Up to Now: Processing and counting individual words
2. Up to Now: Using this as BoW model in information retrieval via vector spaces
3. Now: BoW in text classification via supervised machine learning
4. Problem definition
5. Naive Bayes for topic classification
    1. Multinomial model
    2. Binomial model
6. Next week: evaluation, text classification in practice and NLTK, state of the art, coursework

From IR to text classification: Standing Queries

- IR: ad hoc retrieval, with a ranked output of all documents
- Standing queries: Want to run the query periodically to find **new** items: not ranked but classified as relevant vs. not relevant.
- Example: Google Alerts

Most Frequent Text Classification: Topic Classification

MedLine Article


Brain and Cognition

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

Mesh Subject Categories

- Blood Supply
- Chemistry
- Drug Therapy
- Epidemiology
- Embryology
- ...

Text Classification Definition

Given:

- (Representation of) a document d
- Fixed set of classes (labels, categories) $C = \{c_1, c_2, \ldots, c_j\}$

Determine: Category of d : $\gamma(d) \in C$, where γ is a classification function that maps documents onto classes

Classification Methods I: Manual Classification

- Library of Congress
- Used by the original Yahoo! Directory
- PubMed
- Advantages and Disadvantages?

Classification Methods II: Hand-coded rule-based classifiers

```
comment line                 # Beginning of art topic definition
top-level topic              art ACCRUE
                                 /author = "fsmith"
topic definition modifiers       /date   = "30-Dec-01"
                                 /annotation = "Topic created
                                                by fsmith"
subtopictopic                * 0.70 performing-arts ACCRUE
  evidencetopic              ** 0.50 WORD
  topic definition modifier      /wordtext = ballet
  evidencetopic              ** 0.50 STEM
  topic definition modifier      /wordtext = dance
  evidencetopic              ** 0.50 WORD
  topic definition modifier      /wordtext = opera
  evidencetopic              ** 0.30 WORD
  topic definition modifier      /wordtext = symphony
subtopic                     * 0.70 visual-arts ACCRUE
                             ** 0.50 WORD
                                 /wordtext = painting
                             ** 0.50 WORD
                                 /wordtext = sculpture
subtopic                     * 0.70 film ACCRUE
                             ** 0.50 STEM
                                 /wordtext = film
subtopic                     ** 0.50 motion-picture PHRASE
                             *** 1.00 WORD
                                 /wordtext = motion
                             *** 1.00 WORD
                                 /wordtext = picture
                             ** 0.50 STEM
                                 /wordtext = movie
subtopic                     * 0.50 video ACCRUE
                             ** 0.50 STEM
                                 /wordtext = video
                             ** 0.50 STEM
                                 /wordtext = vcr
                             # End of art topic
```

- Verity (bought by Autonomy, bought by HP `http://www.autonomy.com/technology/`)
- Maintenance issues
- Hand-weighting of terms

Classification Methods III: Supervised ML

Given:

- A (test) document d
- A fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
- A training set D of documents each with a label in C $(d_1, c_1), \ldots (d_m, c_m)$

Determine:

- A learning method which will enable us to learn a classifier γ
- For a test document d, we assign it the class $\gamma(d) \in C$

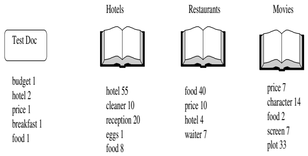
BOW for text classification: intuition

About hotels, restaurants or movies?

A good budget hotel' .... Price includes breakfast with really nice food. Rooms are modern and of a reasonable size. The centre of Leeds is about a 15 min walk at the most. Hotel has bar area.

budget	1
hotel	2
price	1
rooms	1
breakfast	1
food	1
...	...

BOW in a supervised machine learning framework



Naive Bayes: Multinomial Model

Intuition

Use a BoW model with word counts and Bayes rule

For a document d, put into most likely class $c \in C$.

$$
\begin{aligned}
C_{MAP} &= \operatorname*{argmax}_{c \in C} P(c|d) && (1)\\
&= \operatorname*{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} && (2)\\
&= \operatorname*{argmax}_{c \in C} P(d|c)P(c) && (3)\\
&= \operatorname*{argmax}_{c \in C} P(w_1, w_2, \dots w_{n_d}|c)P(c) && (4)\\
&= \operatorname*{argmax}_{c \in C} P(X_1 = w_1, X_2 = w_2, \dots, X_{n_d} = w_{n_d}|c)P(c) && (5)
\end{aligned}
$$

MAP = maximum a posteriori; w_1 word in vocabulary present in position 1 in the document

Naive Bayes: Independence Assumptions

$$P(X_1 = w_1, X_2 = w_2, \dots X_{n_d} = w_{n_d}|c)$$

- Conditional Independence: Assume that the words are conditionally independent given class c.

$$
\begin{aligned}
P(X_1 = w_1, \dots, X_{n_d} = w_{n_d}|c) &= P(X_1 = w_1|c) \cdot P(X_2 = w_2|c)\\
&\quad \cdot \dots \cdot P(X_{n_d} = w_{n_d}|c)
\end{aligned}
$$

- Bag of Words assumption: Assume position of words does not matter

$$P(X_k = w|c) = P(X_l = w|c)$$

for all classes c, all positions l, k and all words w.

Learning the multinomial Naive Bayes classifier

1. From training corpus extract vocabulary
2. Calculate $P(c)$ for all c from training corpus:

$$P(c) = \frac{\#docs\ of\ class\ c}{total\ \#\ of\ docs\ in\ training}$$

3. Calculate $P(w_i|c)$ for all w_i in vocabulary and all c:
   1. Concatenate all training documents of class c into one large doc
   2. $$P(w_i|c) = \frac{n_i^c}{n^c}$$

      where n_i^c is the frequency of word w_i in the large doc and n^c is the length of the large document.

One problem remains: smoothing

Avoid zeros:

$$P(w_i|c) = \frac{n_i^c + 1}{n^c + |V|}$$

where $|V|$ is vocabulary size

Using multinomial NB in testing

$$C_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i \in positions} P(w_i|c)$$

where we range over all positions in the testing document

Multinomial Naive Bayes: a worked example

	docID	words in doc	in c=China?
Training set	1	Chinese Bejing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Testing	5	Chinese Chinese Chinese Tokyo Japan	?

Learning phase: extract vocab, then learn the priors $P(c)$ and the conditional probs $p(w|c)$ from the training set

$p(China = yes) = \frac{3}{4}$, $p(China = no) = \frac{1}{4}$

$p(Chinese|China = yes) = \frac{5+1}{8+6} = \frac{3}{7}$
$p(Tokyo|China = yes) = \frac{0+1}{8+6} = \frac{1}{14}$
$p(Japan|China = yes) = \frac{0+1}{8+6} = \frac{1}{14}$
$p(Chinese|China = no) = \frac{1+1}{3+6} = \frac{2}{9}$
$p(Tokyo|China = no) = \frac{1+1}{3+6} = \frac{2}{9}$
$p(Japan|China = no) = \frac{1+1}{3+6} = \frac{2}{9}$

Multinomial Naive Bayes: A worked Example

Testing phase for testing document 5 *Chinese Chinese Chinese Tokyo Japan*.

$$C_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i \in positions} P(w_i|c) = \operatorname*{argmax}_{c \in C} [\log P(c) + \sum_{i \in positions} \log P(w_i|c)]$$

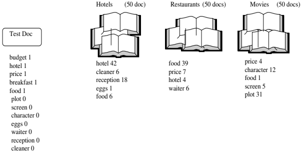
$P(China = yes|d) \propto \frac{3}{4} \cdot \frac{3}{7}^3 \cdot \frac{1}{14} \cdot \frac{1}{14} = 0.0003$
$P(China = no|d) \propto \frac{1}{4} \cdot \frac{2}{9}^3 \cdot \frac{2}{9} \cdot \frac{2}{9} = 0.0001$

Summary of Multinomial Naive Bayes

- Uses conditional independence assumptions
- Uses BoW: ignores position of words in estimating probs
- In training: sees all training documents of one class as one long document.
- In training and testing: cares about frequency of word occurrences!
- Ignores vocabulary words in test document that do not occur

Binomial Naive Bayes

Still uses BoW and conditional independence but cares only about occurrence or non-occurrence of a word, not its frequency!



Binomial Naive Bayes: The derivation

$$\begin{aligned}
C_{NB_{Bi}} &= \operatorname*{argmax}_{c \in C} P(c|d) && (6)\\
&= \operatorname*{argmax}_{c \in C} \frac{P(d|c)P(c)}{P(d)} && (7)\\
&= \operatorname*{argmax}_{c \in C} P(d|c)P(c) && (8)\\
&= \operatorname*{argmax}_{c \in C} P(e_1, \ldots, e_{|v|}|c)P(c) && (9)\\
&= \operatorname*{argmax}_{c \in C} P(c) \prod_{w_i \in V} P(e_i|c) && (10)
\end{aligned}$$

where now e_1 indicates with yes/no whether the word w_1 occurs in the document or not for all vocab items.

Binomial Naive Bayes: A worked example

	docID	words in doc	in c=China?
Training set	1	Chinese Bejing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
Testing	5	Chinese Chinese Chinese Tokyo Japan	?

$p(China = yes) = \frac{3}{4}$, $p(China = no) = \frac{1}{4}$

$p(Chinese|China = yes) = \frac{3+1}{3+2} = \frac{4}{5}$
$p(Japan|China = yes) = p(Tokyo|China = yes) = \frac{0+1}{3+2} = \frac{1}{5}$
$p(Bejing|China = yes) = p(Macao|China = yes) =$
$p(Shanghai|China = yes) = \frac{1+1}{3+2} = \frac{2}{5}$
$p(Chinese|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$
$p(Japan|China = no) = p(Tokyo|China = no) = \frac{1+1}{1+2} = \frac{2}{3}$
$p(Bejing|China = no) = p(Macao|China = no) =$
$p(Shanghai|China = no) = \frac{0+1}{1+2} = \frac{1}{3}$

Binomial Naive Bayes: A worked example

Testing phase for testing document 5 *Chinese Chinese Chinese Tokyo Japan*.

$$C_{NB_{Bi}} = \underset{c \in C}{\text{argmax}}\, P(c) \prod_{w_i \in V} P(e_i|c) = \underset{c \in C}{\text{argmax}}[\log P(c) + \sum_{w_i \in V} \log P(e_i|c)]$$

$$\begin{aligned}
P(China = yes|d) \;&\propto\; P(c) \cdot P(Chinese|China = yes) \cdot P(Japan|China = yes) \\
&\cdot\; P(Tokyo|China = yes) \\
&\cdot\; (1 - P(Bejing|China = yes)) \cdot (1 - P(Shanghai|China = yes)) \\
&\cdot\; (1 - P(Macao|China = yes)) \\
&=\; \frac{3}{4} \cdot \frac{4}{5} \cdot \frac{1}{5} \cdot \frac{1}{5} \cdot (1 - \frac{2}{5}) \cdot (1 - \frac{2}{5}) \cdot (1 - \frac{2}{5}) \\
&=\; 0.005
\end{aligned}$$

Similary we get $P(China = no|d) = 0.022$

Binomial vs Multinomial NB Summary

	multinomial	binomial
random variable	$X = w$ if w occurs at pos	$e_w = 1$ if w occurs in doc
doc rep	sequence of word counts	sequence of zeros and 1s
multiple occurr.	taken into account	ignored
length of docs	good for longer	only good for shorter
number of vocab	can handle more	best for fewer
poss estimate for *the*	$p(X = the|c) = 0.05$	$p(e_{the} = 1|c) = 1.0$

Summary for NB

- Very fast , with low storage requirements
- Robust to irrelevant features (for multinomial)
- Good for domains with many equally important features
- good dependendable baseline for text classification
- If you go beyond topic classification, how do your features need to change? See examples on following slides

Sentiment Classification: Thumbs up or thumbs down?

Original text

A good budget hotel' .... Price includes breakfast. Rooms are modern and of a reasonable size. The centre of Leeds is about a 15 min walk at the most. Hotel has bar area.

Would you use different features than for topic classification?

Text classification: Is this spam?

```
Subject: Conference on the Governments communications strategy
From: Edward Rees
To: Katja Markert

Dear Dr Markert

I hope you wont mind this final reminder about the above seminar,
taking place in Central London on Tuesday, 29th October 2013, but
you dont currently appear to be represented. Please note there is
a  charge for most delegates, although concessionary and
complimentary places are available (subject to terms
and conditions - see below).

The focus:

Proposals and next steps following the Governments
digital strategy - Connectivity, Content and Consumers: Britains digital pl
B; BT Gro
hts Group

ives of c
```

Author identification

- Famous problems: Federalist papers 1787-8.
- Authorship of 12 of the letters in dispute
- 1963: Solved by Mosteller and Wallace using Bayesian methods

Gender/Personality/Age identification

1. The main aim of this article is to propose an exercise in stylistic analysis which can be employed in the teaching of English language. . . . The methods proposed are intended to enable students to obtain insights into aspects of cohesion
2. My aim in this article is to show that given a relevance theoretic approach to utterance interpretation, it is possible to develop a better understanding . . . In this paper I follow Sperber and Wilson's suggestion that . . .

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. Gender, Genre, and Writing Style in Formal Written Texts, Text, volume 23, number 3, pp. 321346

See also: K. Filippova. User demographics and language in an implicit social network. EMNLP 12. Jeju, Korea, July 12-14, 2012.