NLP: Text Classification II

Katja Markert

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

October 30, 2013

Today

1. BoW in text classification via supervised machine learning
2. Problem definition
3. Naive Bayes for topic classification
    1. Multinomial model
    2. Binomial model
4. Now : feature selection, evaluation, extensions with other feature types than words

Features

In the bag of words view of documents

- we used only words as features
- up to now we used all the words in the text OR we left open how to chose the vocabulary that we used as features
- Why might it be better not to use all words in the text?

Answer to “Why Feature Selection”

- Text collections have a large number of features: vocabulary sizes of up to 1m words frequent
- Some classifiers cannot deal with a large number of features: NB can
- Reduces training time and runtime
- Can improve generalisation
  - eliminates noise features
  - Avoides overfitting
- Bernoulli especially sensitive to noise features
- One option is to delete all stop words from the vocabulary

Basic Idea

Feature selection

Compute a utility measure $A(w, c)$ for each word in the vocabulary for a given class c and select the k terms with the highest values of $A(t, c)$

We concentrate on two measures as well as on using two classes only.

Frequency Cutoff points

- Just use the k most common terms
- That does not seem great — includes stop words, excludes many informative words such as *super-entertaining*
- Why does it still work? In practice often almost as good as better methods

An association measure: Mutual Information

Example first: looking at the class `poultry` and the term *export* in the Reuters corpus

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{export} = 1$	$N_{11} = 49$	$N_{10} = 27,652$
$e_t = e_{export} = 0$	$N_{01} = 141$	$N_{00} = 774,106$

$$
\begin{aligned}
MU &= \frac{49}{801,948} \log \frac{801,948 \cdot 49}{(49+27,652)(49+141)} \\
&+ \frac{141}{801,948} \log \frac{801,948 \cdot 141}{(141+774,106)(49+141)} \\
&+ \frac{27,652}{801,948} \log \frac{801,948 \cdot 27,652}{(49+27,652)(27,652+774,106)} \\
&+ \frac{774,107}{801,948} \log \frac{801,948 \cdot 774,106}{(141+774,106)(27,652+774,106)} \\
&= 0.0001105
\end{aligned}
$$

Mutual information: properties

- If term's distribution is the same in the class as in the whole collection, then Mutual info is zero
- Maximum if the term only occurs in the documents of a given class

Mutual information Selection for three Reuters classes

UK class

london 0.192
uk 0.0755
british 0.0596
stg 0.0555
britain 0.0469
plc 0.0357
england 0.0238

sports class

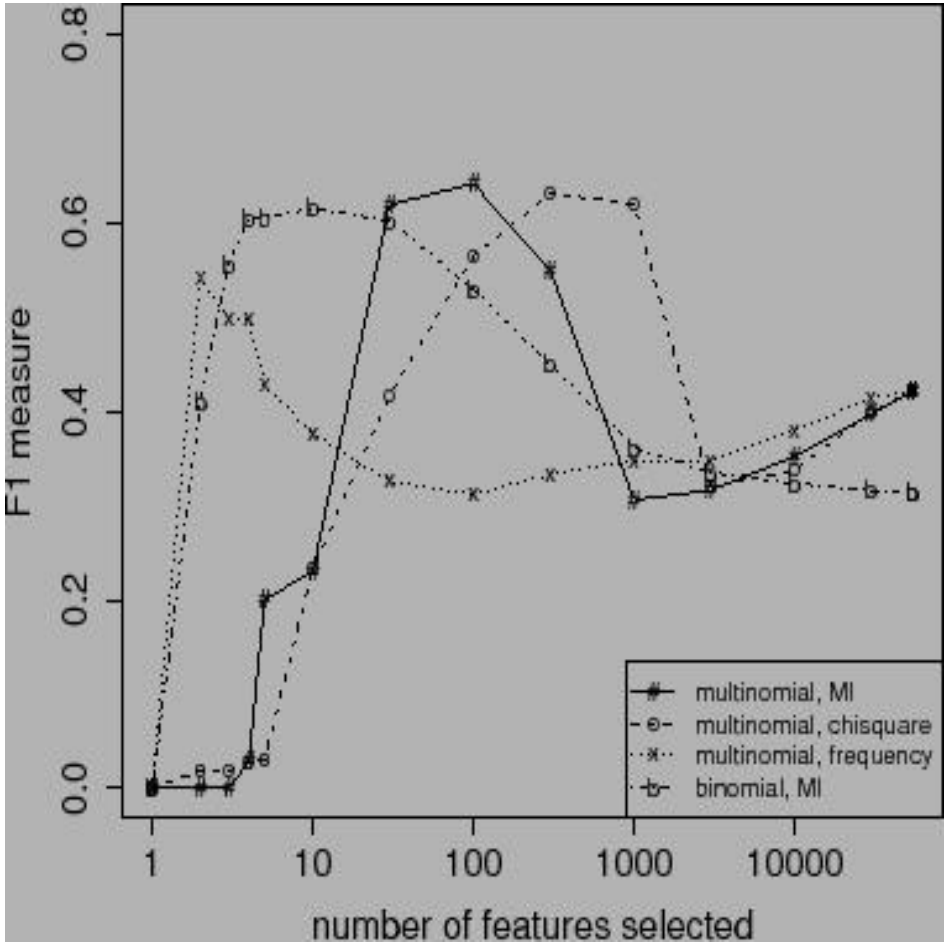
soccer 0.0682
cup 0.0515
match 0.0441
matches 0.0408
played 0.0388
league 0.0386
beat 0.0301

poultry class

poultry 0.0013
meat 0.0008
chicken 0.0006
agriculture 0.0005
avian 0.00004
broiler 0.0003

Effect on performance

For 5 classes yes/no (averaged F-measure over 5 classes) from Manning, Raghavan, Schuetze: Introduction to Information retrieval
100K documents for training and 100K for testing



Evaluation of text classification: Standard contingency table

Same as in IR if you just have two classes

	correct	not correct
selected	tp	fp
not selected	fn	tn

What is precision?
What is recall?
What is F-measure

More than Two Classes: Sets of binary classifiers

- Dealing with any-of or multivalue classifiers
- For each class $c \in C$: build a classifier γ_c to distinguish c from all other casses
- Given test doc d, evaluate it for membership in each class using each γ_c. Then d belongs to any class for which γ_c returns true.

More than Two Classes: Sets of binary classifiers

- One-of classifiers where classes are mutually exclusive
- For each class $c \in C$: build a classifier γ_c to distinguish c from all other casses
- Given test doc d, evaluate it for membership in each class using each γ_c. Then d belongs to the class with the maximum score.

Example Evaluation: Classic Reuters-21578 dataset

- 21,578 docs
- 9603 training, 3299 test articles (ModApte split)
- 118 categories: article can be in more than one category
- Only about 10 categories are large

class	#train	#test	class	#train	#test
earn	2877	1087	trade	369	119
acquisitions	1650	179	interest	347	131
money-fx	538	179	ship	197	89
grain	433	149	wheat	212	71
crude	389	189	corn	182	56

- average document has 1.24 classes

A typical Reuters document


```
<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-
NEWID="798">
<DATE> 2-MAR-1987 16:51:43.42</DATE>
<TOPICS><D>livestock</D><D>hog</D></TOPICS>

<TITLE>AMERICAN PORK CONGRESS KICKS OFF TOMORROW</TITLE>

<DATELINE> CHICAGO, March 2 - </DATELINE><BODY>The American
Pork Congress kicks off tomorrow,
March 3, in Indianapolis with 160 of the nations pork produc
from 44 member states determining industry positions
on a number of issues, according to the National
Pork Producers Council, NPPC.
Delegates to the three day Congress will be considering
26 resolutions concerning various issues, including the futu
direction of farm policy and the tax law
 as it applies to the agriculture sector. The delegates will
endorse concepts of a national PRV (pseudorabies virus)
```

Micro vs. Macro-averaging

You can compute precision and recall per class but then how do you combine multiple performance measures into one?

Macroaveraging

Compute performance for each class, then average

Microaveraging

Collect decisions for all classes, compute contingency table, evaluate

Micro vs. Macroaveraging

Class1

	Truth: y	Truth: n
classifier=y	10	10
classifier =n	10	970

Class2

	Truth: y	Truth: n
classifier=y	90	10
classifier =n	10	890

Micro-averaged table

	Truth: y	Truth: n
classifier=y	100	20
classifier =n	20	1860

- Macroaveraged precision: (0.5+0.9)/2=0.7
- Microaveraged precision: 100/120 = 0.83
- Microaveraged precision is dominated by score on common classes

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:

Features in SpamAssassin for Spam filtering

- Mentions Generic Viagra
- Mentions millions of (dollar) ((dollar) NN,NNN,NNN.NN)
- Phrase:impress...girl
- From: starts with many numbers
- Subject is all capitals
- HTML has a low ratio of text to image area
- Claims you can be removed from the list

`http://spamassassin.apache.org/tests_3_3_x.html`

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.

Summary

In practice you need to

- do some feature selection
- adapt your algorithm to multiple classes
- evaluate carefully
- use non-word features for many text classification tasks