feature selection

March 5, 2017

Overview

Feature selection

- In text classification, we usually represent documents in a high-dimensional space, with each dimension corresponding to a term.
- In this lecture: axis = dimension = word = term = feature
- Many dimensions correspond to rare words.
- Rare words can mislead the classifier.
- Rare misleading features are called noise features.
- Eliminating noise features from the representation increases efficiency and effectiveness of text classification.
- Eliminating features is called feature selection.

Example for a noise feature

- Let's say we're doing text classification for the class *China*.
- Suppose a rare term, say ARACHNOCENTRIC, has no information about *China* . . .
- . . . but all instances of ARACHNOCENTRIC happen to occur in *China* documents in our training set.
- Then we may learn a classifier that incorrectly interprets ARACHNOCENTRIC as evidence for the class *China*.
- Such an incorrect generalization from an accidental property of the training set is called overfitting.
- Feature selection reduces overfitting and improves the accuracy of the classifier.

Basic feature selection algorithm

SelectFeatures($\mathbb{D}, c, k$)
$V \leftarrow$ ExtractVocabulary($\mathbb{D}$)
$L \leftarrow []$
for each $t \in V$
do $A(t, c) \leftarrow$ ComputeFeatureUtility($\mathbb{D}, t, c$)
    Append($L, \langle A(t, c), t \rangle$)
return FeaturesWithLargestValues(L, k)

Basic feature selection algorithm

$\textsc{SelectFeatures}(\mathbb{D}, c, k)$
$V \leftarrow \textsc{ExtractVocabulary}(\mathbb{D})$
$L \leftarrow []$
for each $t \in V$
do $A(t, c) \leftarrow \textsc{ComputeFeatureUtility}(\mathbb{D}, t, c)$
$\quad \textsc{Append}(L, \langle A(t, c), t \rangle)$
return $\textsc{FeaturesWithLargestValues}(L, k)$

How do we compute A, the feature utility?

Different feature selection methods

- A feature selection method is mainly defined by the feature utility measure it employs
- Feature utility measures:
  - Frequency – select the most frequent terms
  - Mutual information – select the terms with the highest mutual information
  - Mutual information is also called information gain in this context.
  - Chi-square (see book)

- Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization.
  In *ICML*, volume 97, pages 412–420, 1997
- Monica Rogati and Yiming Yang. High-performing feature selection for text classification.
  In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 659–661. ACM, 2002
- David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research.
  The Journal of Machine Learning Research, 5:361–397, 2004

feature functions

- These functions capture the intuition that the best terms for c_i are the ones distributed most differently in the sets of positive and negative examples of c_i.
- interpretations of this principle vary across different functions.
- χ^2 is used to measure how the results of an observation differ (i.e. are independent) from the results expected according to an initial hypothesis (lower values indicate lower dependence).

Mutual information

- Compute the feature utility $A(t, c)$ as the mutual information (MI) of term t and class c.
- MI tells us "how much information" the term contains about the class and vice versa.
- For example, if a term's occurrence is independent of the class (same proportion of docs within/without class contain the term), then MI is 0.
- Definition:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{P(U=e_t, C=e_c)}{P(U=e_t)P(C=e_c)}$$

How to compute MI values

- Based on maximum likelihood estimates, the formula we actually use is:

$$\begin{aligned} I(U;C) &= \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} \\ &\quad + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}} \end{aligned}$$

- N_{xy} denote the number of docs that
  - N_{10}: contain t ($e_t = 1$) and are not in c ($e_c = 0$);
  - N_{11}: contain t ($e_t = 1$) and are in c ($e_c = 1$);
  - N_{01}: do not contain t ($e_t = 0$) and are in c ($e_c = 1$);
  - N_{00}: do not contain t ($e_t = 0$) and are not in c ($e_c = 0$);
- $N = N_{00} + N_{01} + N_{10} + N_{11}$.

	Observed			Expected	
	poultry	not poultry	SUM	poultry	no poultry
export	49	27652	27701	6.56	27694.43
no export	141	774106	774247	183.43	774063.56
sum	190	801758	801948		

mutual information intermediate data:

	P(tc)	Obs/Expected	
		7.466090337	
11	6.11012E-05	2.900352965	0.000177215
		0.768656296	
10	0.000175822	-0.379589451	-6.67401E-05
		0.998467671	
01	0.034481039	-0.002212379	-7.62851E-05
		1.000054824	
00	0.965282038	7.90916E-05	7.63457E-05
sum			0.000110536

How to compute MI values (2)

- Alternative way of understanding MI:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U=e_t, C=e_c) \log_2 \frac{N(U=e_t, C=e_c)}{E(U=e_t, C=e_c)}$$

- $N(U=e_t, C=e_c)$ is the count of documents with values e_t and e_c .
- $E(U=e_t, C=e_c)$ is the expected count of documents with values e_t and e_c if we assume that the two random variables are independent.

MI example for *poultry*/EXPORT in Reuters

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

Plug these values into formula:

$$
\begin{aligned}
I(U;C) &= \frac{49}{801{,}948}\log_2\frac{801{,}948\cdot 49}{(49+27{,}652)(49+141)}\\
&\quad+\frac{141}{801{,}948}\log_2\frac{801{,}948\cdot 141}{(141+774{,}106)(49+141)}\\
&\quad+\frac{27{,}652}{801{,}948}\log_2\frac{801{,}948\cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)}\\
&\quad+\frac{774{,}106}{801{,}948}\log_2\frac{801{,}948\cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)}\\
&\approx 0.000105
\end{aligned}
$$

MI feature selection on Reuters

Class: *coffee*

term	MI
COFFEE	0.0111
BAGS	0.0042
GROWERS	0.0025
KG	0.0019
COLOMBIA	0.0018
BRAZIL	0.0016
EXPORT	0.0014
EXPORTERS	0.0013
EXPORTS	0.0013
CROP	0.0012

Class: *sports*

term	MI
SOCCER	0.0681
CUP	0.0515
MATCH	0.0441
MATCHES	0.0408
PLAYED	0.0388
LEAGUE	0.0386
BEAT	0.0301
GAME	0.0299
GAMES	0.0284
TEAM	0.0264

χ^2

$$\chi^2 = \sum_{e_t \in 0,1} \sum_{e_c \in 0,1} \frac{(N_{e_t e_c} - E_{e_t e_c})^2}{E_{e_t e_c}} \tag{1}$$

	Observed			Expected	
	poultry	not poultry	SUM	poultry	no poultry
export	49	27652	27701	6.56	27694.43
no export	141	774106	774247	183.43	774063.56
sum	190	801758	801948		

chi^2 for term *export* and class *poultry*:

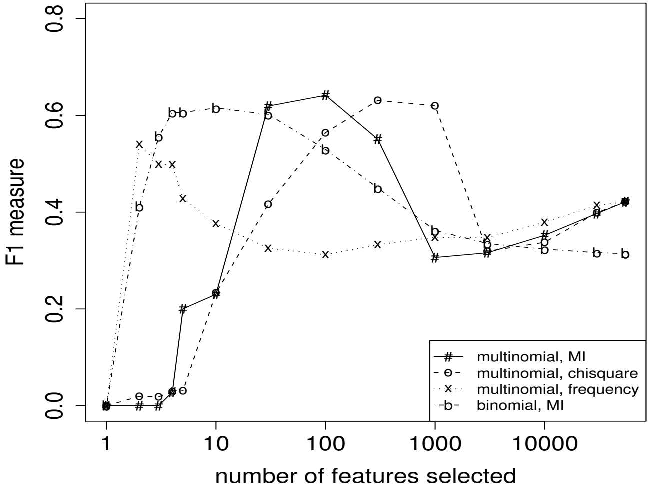
$$\chi^2 = \frac{(49 - 6.56)^2}{6.56} + \frac{(141 - 183)^2}{183} + \ldots \tag{2}$$

$$\approx 274.4 + 9.8 + \ldots \tag{3}$$

$$= 284.2 \tag{4}$$

observed - expected		
	poultry	no poultry
export	42.43699342	(42.44)
not export	(42.43699342)	42.44
square/expected		
	poultry	no poultry
export	274.40143308	0.06502744
not export	9.81753122	0.00232655
sum		284.28631830

Naive Bayes: Effect of feature selection



multinomial = multinomial Naive Bayes
binomial = Bernoulli Naive Bayes

Feature selection for Naive Bayes

- In general, feature selection is necessary for Naive Bayes to get decent performance.
- Also true for many other learning methods in text classification: you need feature selection for optimal performance.

Exercise

- Compute the "export"/POULTRY contingency table for the "Kyoto"/JAPAN in the collection given below.
- Make up a contingency table for which MI is 0 – that is, term and class are independent of each other.

"export"/POULTRY table:

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

Collection:

	docID	words in document	in c = Japan?
training set	1	Kyoto Osaka Taiwan	yes
	2	Japan Kyoto	yes
	3	Taipei Taiwan	no
	4	Macao Taiwan Shanghai	no
	5	London	no

[LYRL04] David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[RY02] Monica Rogati and Yiming Yang. High-performing feature selection for text classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 659–661. ACM, 2002.

[YP97] Yiming Yang and Jan O Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.