



Text Classification and Naïve Bayes

Text Classification: Evaluation



More Than Two Classes: Sets of binary classifiers

- Dealing with **any-of** or **multivalue** classification
 - A document can belong to 0, 1, or >1 classes.
- For each class $c \in C$
 - Build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test doc d ,
 - Evaluate it for membership in each class using each γ_c
 - d belongs to **any** class for which γ_c returns true



More Than Two Classes: Sets of binary classifiers

- One-of or multinomial classification
 - Classes are mutually exclusive: each document in exactly one class
- For each class $c \in C$
 - Build a classifier γ_c to distinguish c from all other classes $c' \in C$
- Given test doc d ,
 - Evaluate it for membership in each class using each γ_c
 - d belongs to the one class with maximum score



Evaluation:

Classic Reuters-21578 Data Set

- Most (over)used data set, 21,578 docs (each 90 types, 200 tokens)
- 9603 training, 3299 test articles (ModApte/Lewis split)
- 118 categories
 - An article can be in more than one category
 - Learn 118 binary category distinctions
- Average document (with at least one category) has 1.24 classes
- Only about 10 out of 118 categories are large

Common categories
(#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) |



Reuters Text Categorization data set (Reuters-21578) document

Sec. 15.2.4

<REUTERS TOPICS="YES" LEWISSPLIT="TRAIN" CGISPLIT="TRAINING-SET" OLDID="12981" 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 producers 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 future direction of farm policy and the tax law as it applies to the agriculture sector. The delegates will also debate whether to endorse concepts of a national PRV (pseudorabies virus) control and eradication program, the NPPC said.

A large trade show, in conjunction with the congress, will feature the latest in technology in all areas of the industry, the NPPC added. Reuter

⁵⁷
</BODY></TEXT></REUTERS>



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.
- **Microaveraging:** Collect decisions for all classes, compute contingency table, evaluate.



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

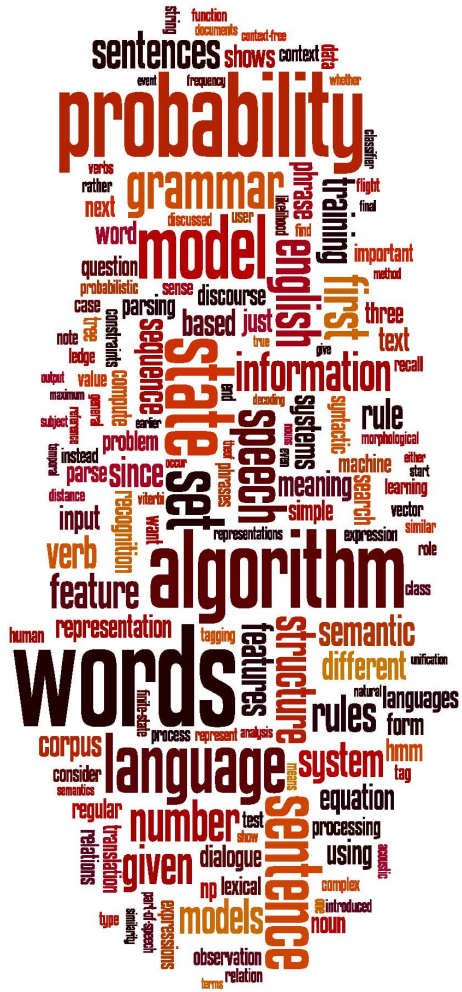
- Metric: P/R/F1 or Accuracy
- Unseen test set
 - avoid overfitting ('tuning to the test set')
 - more conservative estimate of performance
- Cross-validation over multiple splits
 - Handle sampling errors from different datasets
 - Pool results over each split
 - Compute pooled dev set performance

Training Set Dev Test

Training Set Dev Test

Dev Test Training Set

Test Set



Text Classification and Naïve Bayes

Text Classification: Evaluation