

Feature Selection

What is feature selection? Why do it?

1

Feature Selection: Why?

- Text collections have a large number of features
 - 10,000 – 1,000,000 unique words ... and more
- May make using a particular classifier feasible
 - Some classifiers can't deal with 100,000 of features
- Reduces training time
 - Training time for some methods is quadratic or worse in the number of features
- Can improve generalization (performance)
 - Eliminates noise features
 - Avoids overfitting

General idea:

For each term t in vocabulary V and each class c , compute
“utility measure” $A(t,c)$

Select the k terms with highest $A(t,c)$

3

Feature selection via Mutual Information

- In training set, choose k words which best discriminate (give most info on) the categories.
- The Mutual Information between a word, class is:

$$I(t, c) = \sum_{e_t \in \{0,1\}} \sum_{e_c \in \{0,1\}} p(e_t, e_c) \log_2 \frac{p(e_t, e_c)}{p(e_t)p(e_c)}$$

for each term t and each category c ,

where $e_t = 1$ if term t is in document, 0 otherwise
and $e_c = 1$ if the document is class c , 0 otherwise

Example

- Suppose we have 100 documents, 60 of which contain the term “dollar” and 50 of which don’t. Of the 60 that contain “dollar”, 45 are class “spam”. Of the 50 that do not contain “dollar”, 20 are spam.

What is $I(\text{“dollar”, spam})$?

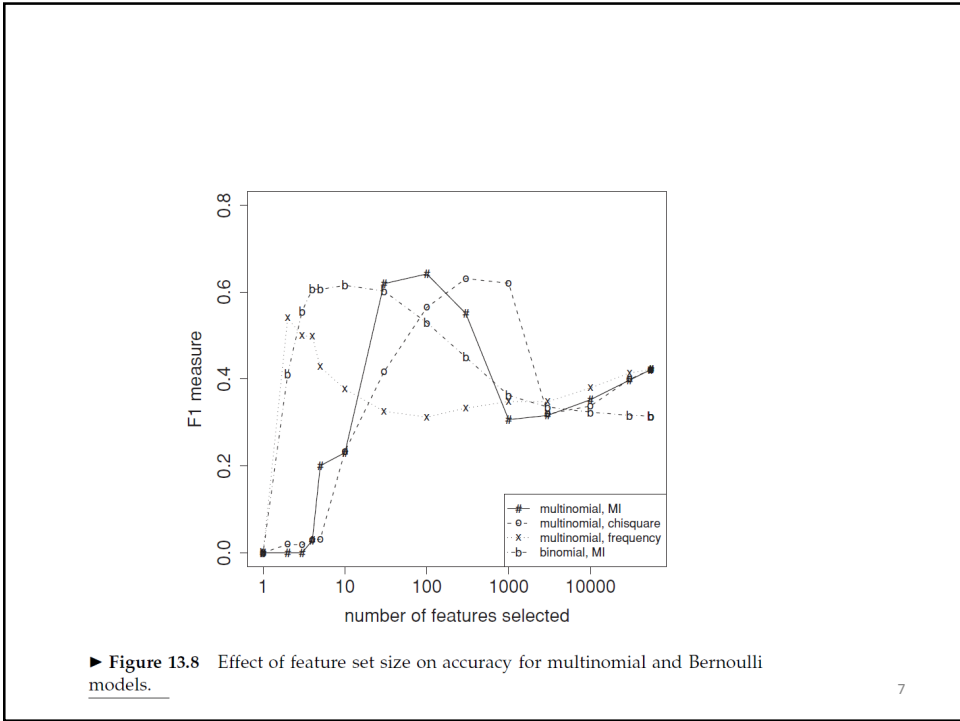
$$I(\text{“dollar”, spam}) = \sum_{e_{\text{dollar}} \in \{0,1\}} \sum_{e_{\text{spam}} \in \{0,1\}} p(e_{\text{dollar}}, e_{\text{spam}}) \log_2 \frac{p(e_{\text{dollar}}, e_{\text{spam}})}{p(e_{\text{dollar}})p(e_{\text{spam}})}$$

5

<i>UK</i>		<i>China</i>		<i>poultry</i>	
london	0.1925	china	0.0997	poultry	0.0013
uk	0.0755	chinese	0.0523	meat	0.0008
british	0.0596	beijing	0.0444	chicken	0.0006
stg	0.0555	yuan	0.0344	agriculture	0.0005
britain	0.0469	shanghai	0.0292	avian	0.0004
plc	0.0357	hong	0.0198	broiler	0.0003
england	0.0238	kong	0.0195	veterinary	0.0003
pence	0.0212	xinhua	0.0155	birds	0.0003
pounds	0.0149	province	0.0117	inspection	0.0003
english	0.0126	taiwan	0.0108	pathogenic	0.0003
<i>coffee</i>		<i>elections</i>		<i>sports</i>	
coffee	0.0111	election	0.0519	soccer	0.0681
bags	0.0042	elections	0.0342	cup	0.0515
growers	0.0025	polls	0.0339	match	0.0441
kg	0.0019	voters	0.0315	matches	0.0408
colombia	0.0018	party	0.0303	played	0.0388
brazil	0.0016	vote	0.0299	league	0.0386
export	0.0014	poll	0.0225	beat	0.0301
exporters	0.0013	candidate	0.0202	game	0.0299
exports	0.0013	campaign	0.0202	games	0.0284
crop	0.0012	democratic	0.0198	team	0.0264

► **Figure 13.7** Features with high mutual information scores for six Reuters-RCV1 classes.

6



7

Feature selection via χ^2 statistic (CHI)

- χ^2 is interested in $(f_o - f_e)^2 / f_e$ summed over all table entries: is the observed number what you'd expect given the marginals?

$$\chi^2(j, a) = \sum (O - E)^2 / E = (2 - .25)^2 / .25 + (3 - 4.75)^2 / 4.75 + (500 - 502)^2 / 502 + (9500 - 9498)^2 / 9498 = 12.9 \quad (p < .001)$$

- The null hypothesis is rejected with confidence .999,
- since $12.9 > 10.83$ (the value for .999 confidence).

	Term = jaguar	Term ≠ jaguar	
Class = auto	2 (0.25)	500 (502)expected: f_e
Class ≠ auto	3 (4.75)	9500 (9498)observed: f_o

Evaluating Categorization

- Evaluation must be done on test data that are independent of the training data (usually a disjoint set of instances).
 - Sometimes use cross-validation (averaging results over multiple training and test splits of the overall data)
- It's easy to get good performance on a test set that was available to the learner during training (e.g., just memorize the test set).
- Measures: precision, recall, F1, classification accuracy
- *Classification accuracy*: c/n where n is the total number of test instances and c is the number of test instances correctly classified by the system.
 - Adequate if one class per document
 - Otherwise F measure for each class