Text Categorization: Evaluation

Part 1

ChengXiang "Cheng" Zhai
Department of Computer Science
University of Illinois at Urbana-Champaign

Overview

- What is text categorization?

- Why text categorization?

- How to do text categorization?
  - Generative probabilistic models
  - Discriminative approaches

- **How to evaluate categorization results?**

General Evaluation Methodology

- Have humans to create a test collection where every document is tagged with the desired categories ("ground truth")
- Generate categorization results using a system on the test collection
- Compare the system categorization decisions with the human-made categorization decisions and quantify their similarity (or equivalently difference)
  - The higher the similarity is, the better the results are
  - Similarity can be measured from different perspectives to understand the quality of results in detail (e.g., which category performs better?)
  - In general, different categorization mistakes may have a different cost that inevitably depends on specific applications, but it is okay not to consider such a cost variation for **relative comparison of methods**

Classification Accuracy (Percentage of Correct Decisions)

	c_1	c_2	c_3	...	c_k
d_1	y(+)	y(-)	n(+)		n(+)
d_2	y(-)	n(+)	y(+)		n(+)
d_3	n(+)	n(+)	y(+)		n(+)
...					
d_N	...	...			

+/- human answer
(+= correct; - =incorrect)

y/n   system result
 (y=yes; n=no)

Classification Accuracy = $\dfrac{\textbf{Total number of correct decisions}}{\textbf{Total number of decisions made}}$

$$= \frac{\mathrm{count}(y(+)) + \mathrm{count}(n(-))}{kN}$$

4

Problems with Classification Accuracy

- Some decision errors are more serious than others
  - It may be more important to get the decisions right on some documents than others
  - It may be more important to get the decisions right on some categories than others
  - E.g., spam filtering: missing a legitimate email costs more than letting a spam go
- Problem with imbalanced test set
  - Skewed test set: 98% in category 1; 2% in category 2
  - Strong baseline: put all instances in category 1 ➜ 98% accuracy!

Per-Document Evaluation

	c_1	c_2	c_3	...	c_k
d_1	y(+)	y(-)	n(+)		n(+)
d_2	y(-)	n(+)	y(+)		n(+)
d_3	n(+)	n(+)	y(+)		n(+)

How good are the decisions on d_i?

When the system says "yes," how many are correct?

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

	System ("y")	System ("n")
Human (+)	True Positives TP	False Negatives FN
Human (-)	False Positives FP	True Negatives TN

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Does the doc have all the categories it should have?

Per-Category Evaluation

	c_1	c_2	c_3	...	c_k
d_1	y(+)	y(-)	n(+)		n(+)
d_2	y(-)	n(+)	y(+)		n(+)
d_3	n(+)	n(+)	y(+)		n(-)

How good are the decisions on c_i?

When the system says "yes," how many are correct?

$$Precision = \frac{TP}{TP + FP}$$

	System ("y")	System ("n")
Human (+)	True Positives TP	False Negatives FN
Human (-)	False Positives FP	True Negatives TN

$$Recall = \frac{TP}{TP + FN}$$

Has the category been assigned to all the docs of this category?

Combine Precision and Recall: F-Measure

$$F_\beta = \cfrac{1}{\frac{\beta^2}{\beta^2+1}\frac{1}{R} + \frac{1}{\beta^2+1}\frac{1}{P}} = \frac{(\beta^2+1)P*R}{\beta^2 P + R}$$

$$F_1 = \frac{2PR}{P+R}$$

P: precision

R: recall

β: parameter (often set to 1)

Why not 0.5*P+0.5*R?

What is R if the system says "y" for all category-doc pairs?