CSCI 4152/6509 — Natural Language Processing

2-Oct-2009

Lecture 10: Text Classification

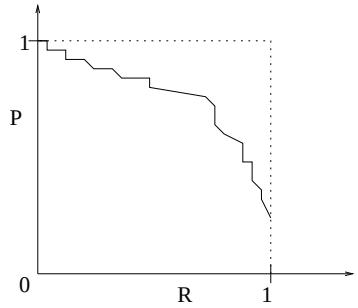
Room: FASS 2176
Time: 11:35 – 12:25

Previous Lecture

- aside: Lucene, handout: NL Principles in Perl;
- Typical IR system architecture,
- steps in document and query processing in IR,
- vector space model,
- tfidf term weighting formula,
- cosine similarity measure,
- term-by-document matrix,
- reducing the number of dimensions,
- Latent Semantic Analysis,
- IR evaluation

IR Evaluation

- Precision and Recall
- Precision-Recall Curve



- Interpolated Precision: $\mathit{IntPrec}(r) = \max_{i \geq r} \mathit{Prec}(i)$
- F-Measure

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

Text Mining

The two main tasks in Text Mining are:

- Text Classification, and
- Text Clustering

6.2 Text Classification

- It is also known as Text Categorization.
- additional reading: Manning and Schütze, Ch 16: Text Categorization
- Problem definition
- An example of supervised learning

Types of Text Classification

Some types of text classification:

- spam detection and e-mail classification
- encoding and language identification
- sentiment classification
- authorship attribution and plagiarism detection
- automatic essay grading
- topic categorization

More specialized: dementia detection using spontaneous speech

Evaluation Measures for Text Classification

- contingency table, or confusion matrix; all-class table, or per class:
  Yes – in class; No — not in class

	Yes is correct	No is correct
Yes assigned	a	b
No assigned	c	d

- accuracy ($\frac{a+d}{a+b+c+d}$), precision ($\frac{a}{a+b}$), recall ($\frac{a}{a+c}$), fallout ($\frac{b}{b+d}$), F-measure

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- $\beta = 1 \Rightarrow$ Precision and Recall treated equally, $\beta > 1 \Rightarrow$ Recall given higher weight, and vice versa.
- macro-averaging (equal weight to each class) and micro-averaging (equal weight to each object)
  (2×2 contingency tables vs. one large contingency table)

Evaluation Methods for Classification

- General issues in classification: overfitting and underfitting
- Example with polynomial-based function learning
- Evaluation methods in classification:
  1. training error
  2. train and test
  3. n-fold cross-validation

Training Error. The classifier is trained on a training data set and also evaluated on the same data set. It is a good idea to get this result, although it is obviously *biased* towards the training data. This evaluation can detect underfitting but not overfitting of the training data.

Train and test. The data is divided into two parts: training and testing part. The split is usually 90% for training and 10% for testing, but sometimes 2/3 of data is used for training and 1/3 for testing. This is an unbiased evaluation, which can detect underfitting as well as overfitting. To be sure that the evaluation is unbiased, it is important not to use testing data in any way, even to glance at it, if it may influence our decisions regarding classifier construction. With some methodologically generic methods, this is not an issue.