Natural Language Processing CSCI 4152/6509 — Lecture 7 Elements of Information Retrieval and Text Mining

Instructors: Vlado Keselj
Time and date: 16:05 – 17:25, 25-Sep-2024
Location: Carleton Tupper Building Theatre C

Previous Lecture

- Elements of Morphology (continued):
  - Lemmatization, Morphological Processes
- Word Counting and Zipf's Law
- N-grams definition
- Extracting and Analyzing n-grams in Perl
- Elements of Information Retrieval
  - Vector space model

Side Note: Lucene and IR Book

- Lucene search engine
- http://lucene.apache.org
- Open-source, written in Java
- Uses the vector space model
- Another interesting link: Introduction to IR on-line book covers well text classification:

  http://nlp.stanford.edu/IR-book/html/htmledition/irbook.html

IR Evaluation: Precision and Recall

- **Precision** is the percentage of true positives out of all returned documents; i.e.,

$$P = \frac{\mathit{TP}}{\mathit{TP} + \mathit{FP}}$$

- **Recall** is the percentage of true positives out of all relevant documents in the collection; i.e.,

$$R = \frac{\mathit{TP}}{\mathit{TP} + \mathit{FN}}$$

Precision and Recall: Venn Diagram

F-measure

- **F-measure** is a weighted harmonic mean between Precision and Recall:

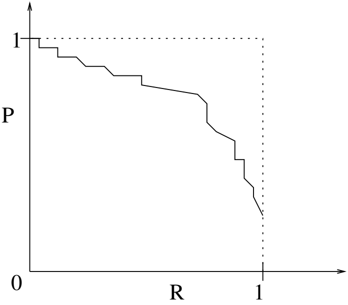
$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- We usually set $\beta = 1$, in which case we have:

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

Recall-Precision Curve

- A more appropriate way to evaluate a ranked list of relevant documents is the Recall-Precision Curve
- Connects (recall, precision) points for the sets of 1, 2, ...most relevant documents on the list
- It typically looks as follows:



Recall-Precision Curve Example

Results returned by a search engine (8 rel.doc.total):

1. relevant
2. relevant
3. relevant
4. not relevant
5. relevant
6. not relevant
7. relevant
8. not relevant
9. not relevant
10. relevant
11. not relevant
12. not relevant

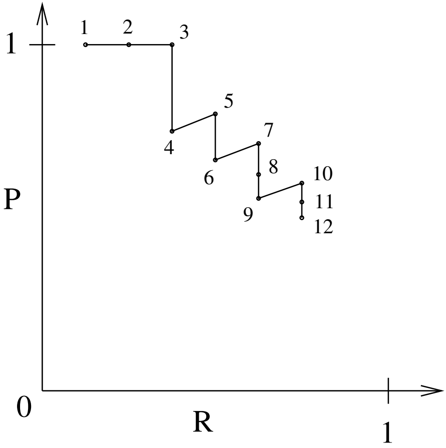
Task 1: Precision, Recall and F-measure

- Assuming that the total number of relevant documents in the collection is 8, calculate precision, recall, and F-measure ($\beta = 1$) for the returned 12 results.

Task 2: Recall-Precision Curve

- Task: Draw the recall-precision curve for these results
- First step: Form sets of n initial documents, and look at their relevance:
  - Set 1: $\{R\}$ $(R = 0.125, P = 1)$
  - Set 2: $\{R, R\}$ $(R = 0.25, P = 1)$
  - Set 3: $\{R, R, R\}$, $(R = 0.375, P = 1)$
  - Set 4: $\{R, R, R, NR\}$, $(R = 0.375, P = 0.75)$
  - Set 5: $\{R, R, R, NR, R\}$, $(R = 0.5, P = 0.8)$
  - ...etc.

Recall-Precision Curve



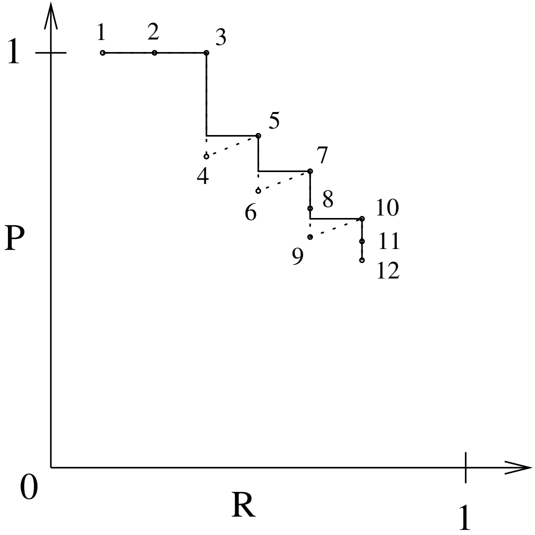
Task 3: Interpolated Recall-Precision Curve

- Task: Draw interpolated Recall-Precision curve
- Formula:

$$\mathit{IntPrec}(r) = \max_{k,R(k)\geq r} P(k)$$

- Based on the previous Task:
  $0 \leq r \leq R_4 = \frac{3}{8} = 0.375 \Rightarrow \mathit{IntPrec}(r) = 1$
  $R_4 < r \leq R_6 = \frac{4}{8} = 0.5 \Rightarrow \mathit{IntPrec}(r) = 0.8$
  $R_6 < r \leq R_9 = \frac{5}{8} = 0.625 \Rightarrow \mathit{IntPrec}(r) = 5/7 \approx 0.714285714$
  $R_9 < r \leq R_{12} = \frac{6}{8} = 0.75 \Rightarrow \mathit{IntPrec}(r) = 0.6$

Interpolated Recall-Precision Curve



Interpolated R-P Curve at 11 Standard Levels

Some Other Similar Measures

- Fallout

$$Fallout = \frac{FP}{FP + TN}$$

- Specificity

$$Specificity = \frac{TN}{TN + FP}$$

- Sensitivity

$$Sensitivity = \frac{TP}{TP + FN} \quad (= R)$$

- Sensitivity and Specificity: useful in classification and contexts such as medical tests

Some Text Mining Tasks

- Text Classification
- Text Clustering
- Information Extraction
- And some new and less prominent tasks:
  - Text Visualization
  - Filtering tasks, Event Detection
  - Terminology Extraction

Text Classification

- It is also known as Text Categorization.
- Additional reading: Manning and Schütze, Ch 16: Text Categorization
- Problem definition:
  Classify a document into a class (category) of documents
- Typical approach:
  Use of Machine Learning to learn classification model from previously labeled documents
- An example of supervised learning

Types of Text Classification

- topic categorization
- sentiment classification
- authorship attribution and plagiarism detection
- authorship profiling (e.g., age and gender detection)
- spam detection and e-mail classification
- encoding and language identification
- automatic essay grading

More specialized example: dementia detection using spontaneous speech

Creating Text Classifiers

- Can be created manually
  - typically rule-based classifier
  - example: detect or count occurrences of some words, phrases, or strings
- Another approach: make programs that *learn* to classify
  - In other words, classifiers are generated based on labeled data
  - supervised learning

Evaluation Measures for Text Classification

- Contingency table (confusion matrix) and Accuracy
- Example (classes A, B, and C):

		Gold standard			
		A	B	C	
Model classification	A	5	1	1	7
	B	3	10	2	15
	C	0	2	10	12
		8	13	13	34

- Accuracy: percentage of correct classifications; in the example, $= 25/34 \approx 0.7353 = 73.53\%$

Per class: Precision, Recall, and F-measure

- For each class: Yes = in class, No = not in class

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

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

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

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

Example: Classification Results

		Gold standard			
		A1	A2	A3	
System response	A1	5	1	1	7
	A2	3	10	2	15
	A3	0	2	10	12
		8	13	13	34

Or, we can create contingency tables for each class separately:

	Gold standard		
	A1	not A1	
A1	5	2	7
not A1	3	24	27
	8	26	34

	Gold standard		
	A2	not A2	
A2	10	5	15
not A2	3	16	19
	13	21	34

	Gold standard		
	A3	not A3	
A3	10	2	12
not A3	3	19	22
	13	21	34

The overall accuracy can be calculated using the overall table;

$$\mathit{Accuracy} = \frac{5+10+10}{34}$$

Per-class precisions are:

$$P_{A1} = \frac{5}{7} \quad P_{A2} = \frac{10}{15} \quad P_{A3} = \frac{10}{12}$$

Per-class recalls are:

$$R_{A1} = \frac{5}{8} \quad R_{A2} = \frac{10}{13} \quad R_{A3} = \frac{10}{13}$$

Macro-averaged precision, recall, and F-measure are:

$$P_{macro} = \frac{5/7 + 10/15 + 10/12}{3} \quad R_{macro} = \frac{5/8 + 10/13 + 10/13}{3}$$

$$F_{macro} = \frac{2 \cdot P_{macro} \cdot R_{macro}}{P_{macro} + R_{macro}}$$

To calculate micro-averaged precision, recall, and F-measure, we calculate cumulative per-class table:

	Gold standard		
	A	not A	
A	25	9	34
not A	9	59	68
	34	68	102

and then we calculate the micro-averaged measures:

$$P_{micro} = \frac{25}{34} \quad R_{micro} = \frac{25}{34} \quad F_{micro} = \frac{2 \cdot P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} = \frac{25}{34}$$

Evaluation Methods for Classification

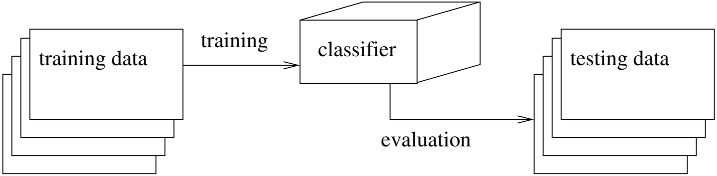
- General issues in classification
  - Underfitting and Overfitting
- Example with polynomial-based function learning
  - Underfitting and Overfitting

Evaluation Methods for Text Classifiers

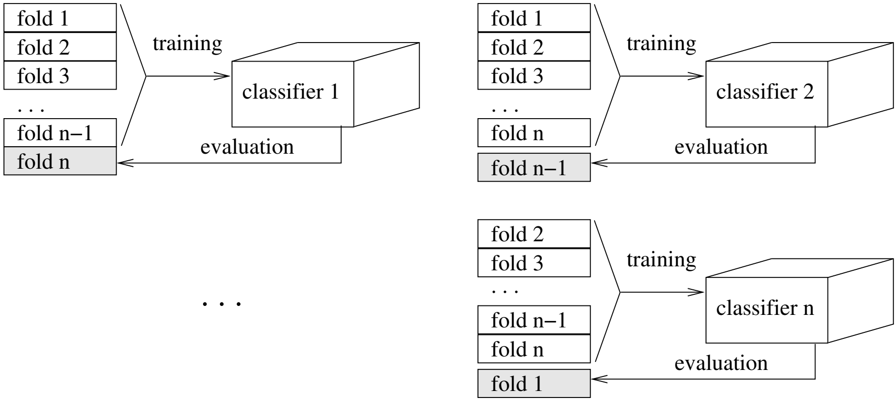
- Training Error
- Train and Test
- N-fold Cross-validation

Train and Test

- Labeled data is divided into training and testing data
- Typically training data size : testing data size $= 9 : 1$, sometimes $2 : 1$



N-fold Cross-Validation



Text Clustering

- Text clustering is an interesting text mining task
- It is relevant to the course and a clustering task can be a project topic
- Since it is covered in some other courses, we will not cover it in much detail here
- Some notes are provided for your information

Similarity-based Text Classification

- Aggregate training text for each class into a profile
- Aggregate testing text into another profile
- Classify according to profile similarity
- If a profile is a vector, we can use different similarity measures; e.g.,
  - cosine similarity,
  - Euclidean similarity, or
  - some other type of vector similarity