Exam INF5830, 2011, Some solutions

Exercise 1

Lee is experimenting with word sense disambiguation. He uses the hard-data from the famous hard-line-serve-corpus, and reserves 433 (1 out of 10) occurrences for testing and uses the rest for training. This is the result from one of his runs.

		Correct class		
		Hard 1	Hard 2	Hard 3
	Assigned Hard 1	341	28	23
	Assigned Hard 2	6	15	0
	Assigned Hard 3	2	1	17

For each occurrence of hard, Lee is interested in whether it is classified correctly or not. What is the accuracy of the classifier in this sense (i.e., micro-averaged accuracy).

		Correct class			
		Hard 1	Hard 2	Hard 3	
	Assigned Hard 1	341	28	23	392
	Assigned Hard 2	6	15	0	21
	Assigned Hard 3	2	1	17	20
		349	44	40	433

Accuracy = (341+15+17)/433 = 373/433 = 0.861

(Baseline: 0.806)

Suppose we are only interested in how well the classifier is doing as a classifier for Sense3. This corresponds to merging Sense1 and Sense2. What is the accuracy, recall and precision for the classifier for Sense3? You may present the answers as fractions.

		Correct class		
		Hard 1 + Hard 2	Hard 3	
	Assigned Hard 1 + Assigned Hard 2	341+28+6+15=390	23+0=23	392+21=413
	Assigned Hard 3	2+1=3	17	20
		349+44=393	40	433

Accuracy: (390+17)/433=407/433=0.940

Precision: 17/20=0.85

Recall 17/40=0.425

(Comments: f-score was not included since it is harder to calculate without a calculator.
If we do similar calculations for sense 1 and sense 2 we may take the averages for the three and get macroaveraged accuracy, precision, recall. It makes even more sense to consider one class at a time for an any-of-classifier (compared to a one-of).

Exercise 4

Kim has constructed a system for classifying textual entailment. She has read that the best classifiers have 0.72 accuracy when tested on test material which is evenly split between entailment and non-entailment examples, and this is the baseline she wants to beat. Since Kim is very careful and avoids seeing the test items herself, she has got her fellow students to construct a test base of 50 examples of entailment and 50 examples of non-entailment, annotate the examples and shuffle them. We may assume that these test examples can be considered a random sample. When she runs her classifier on the test set, it classifies 75 out of the 100 items correctly. At first Kim gets very excited and wants to publish her results immediately. But then she remembers vaguely something about statistical significance and comes to you for help.

a) Is Kim's result statistically significantly better than the baseline at the 0.05 level?
Since we do not use computers you may use 0.2 for 0.72×(1 – 0.72) (instead of 0.2016). You may also use 0.45 for the square root of 0.72×(1 – 0.72) (instead of 0.4490).

The population consists of all candidates for entailment/non-entailment. We are interested in the proportion of these that are classified correctly by the baseline classifier. This proportion is known to be 0.72. This means that if we choose a random object the chance it is classified correctly by the baseline classifier is p=0.72. If we pick random samples of *n* individuals, the probability of classifying *k* many correctly follows a binomial distribution. When *n* is reasonably large and 0.1<p<0.9 we may approximate by the normal distribution We also know the true standard deviation σ, where σ^2=p(1-p).

$$z = \frac{\bar{x}-\mu}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} = \frac{0.75-0.72}{\frac{\sqrt{0.72(1-0.72)}}{\sqrt{100}}} = \frac{0.03}{\frac{0.45}{10}} = \frac{30}{45} = \frac{2}{3}$$

This is not significant. The small table tells we need a z-value of 1.645.

b) Kim is a very careful student and thinks the 0.05 level is too little. She wants a result which is statistically significant at the 0.01 level. She realizes that she will need a larger test corpus for this and that she must bribe her fellow students into marking up some more examples. But how many test items is necessary to show that a raise in performance from 0.72 to 0.75 (i.e. 3 per cent points) is statistically significant at the 0.01 level?

According to the table, the required z-value is 2.326. We put this into the formula and get

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$z\sigma = (\bar{x} - \mu)\sqrt{n}$$

$$n = \left(\frac{z\sigma}{\bar{x} - \mu}\right)^2$$

$$n = \left(\frac{2.326 \times 0.45}{0.75 - 0.72}\right)^2 = \left(\frac{2.326 \times 45}{3}\right)^2 = (2.326 \times 15)^2 = 34.89^2 < 35^2 = 1225$$

Hence 1225 items should suffice.