INF5830 – 2013, Obligatory assignments, set 2

Part A

To be delivered by Oct. 16, 18:00 (6 p.m.)

Observe, this is only the first part of obligatory assignment 2. There is also a part B. Both parts should be completed!

1 Classification, accuracy and interval estimation

a) The starting point is chapter 6 in the NLTK book. Construct the name classifier based on suffix features, but split the corpus in 1000 items for testing, 1000 items for development and the rest for training. Split the test set in 10 disjoint sets, each containing 100 items. Evaluate the classifier with each of the 10 test sets and report the accuracy for each test set. Do you get the same accuracy for each test set? Is there any lessons to be learnt from this?

b) We may consider the accuracy a random variable, and we have ten observations of this variable. We may use this for estimating the true accuracy. What is the mean, variance and standard deviation for the 10 observations? From the ten observations, estimate an interval for the true accuracy with a 95% confidence level.

c) Say you only had the result from the first test set, call it test set ONE. An alternative way for estimating the true accuracy could be to do it theoretically from this test set alone. The accuracy could be considered the result of 100 Bernoulli trials. Each trial has two possible outcomes, 1 for correct, and 0 for incorrect. Estimate an interval for the true accuracy from the first test set alone in this way, and explain how you proceed.

d) Now, unify the 10 test sets to one big test set containing 1000 items. Call it BIG. What is the accuracy when testing towards this set? Estimate an interval for the true accuracy from this test set with a 95% confidence level. Do you get the same result as when you used test set ONE? Are any of the results from ONE or BIG the same as when you estimated from the 10 observation in (a)? If not, is the result from (a) closest to ONE or to BIG? Why do you think it is so?

2 Evaluation and hypothesis testing

a) You are constructing a parser. You are comparing your parser to a reference parser. While the reference parser parses 90% of all sentences correctly, your parser parses 91.25% of all sentences correctly. You are very pleased with the results, but you remember something about chance and significance and want to make sure that this result is not a result of pure luck. You established the numbers as follows. You parsed 1600 sentences from a reference corpus with the reference parser, parser A, and 1440 of them were correctly parsed. Then you parsed 1600 sentences from the reference corpus with your own parser, parser B, and 1460 of them were parsed correctly. What is the p-value you get when you compare the two with the two-sample t-test?

$$t = \frac{\overline{x}_B - \overline{x}_A}{\sqrt{\frac{s_A^2}{N} + \frac{s_B^2}{N}}}$$

Would you conclude that parser B is better than parser A?

b) Luckily, you have some more information. You did not test the two parsers on different test sets but on the same set and you kept records for each sentence. Some sentences seem hard for both parsers and the joint results may be summarized as follows.

- There are 1415 sentences where both parsers succeeded.
- There are 25 sentences where A succeeded and B failed.
- There are 45 sentences where B succeeded and A failed.
- There are 115 sentences where both parsers failed.

We may define a new random variable $X_C = X_B - X_A$. The null hypothesis is that $\mu_C = 0$. We may perform a T-test of the null hypothesis. What is the result of the test? Can you now conclude that parser B is better than parser A?

c) What have you learned from this exercise?

To be continued in part B