Coursework **Spring 2001**
Text Processing: Inductive Techniques and Applications (150TP)
EECS Department, Tufts University **Roni Khardon**

Pencil&Paper Assignment 3

This assignment is due by Monday, April 23th (in class).

In practical exercise 2 you made experiments with several learning algorithms and reported accuracy results. In this exercise you should derive confidence intervals for some of these results. In particular we will look at he 3-fold cross validation experiment with winnow. We will base the confidence intervals on the following.

1. The central limit theorem: if $x_1, \ldots, x_n$ are distributed IID with expectation and standard deviation μ, σ then (for large n) $\bar{x} = \frac{1}{n}\sum x_i$ is distributed approximatly with a normal distribution $N(\mu, \sigma^2/n)$

2. If $z_1, \ldots, z_n$ are distributed $N(\mu, \sigma)$ then with confidence $(1-\alpha)$ we have

$$\mu \in \bar{z} \pm \frac{\sigma}{\sqrt{n}} z(1 - \frac{\alpha}{2}) \tag{1}$$

   where $z()$ is the quantile of the normal distribution.

3. If σ is not known we can try to approximate this by estimating σ, or bounding it.

4. Alternatively, we can use the next interval: with confidence $(1-\alpha)$ we have

$$\mu \in \bar{z} \pm \frac{1}{\sqrt{n-1}} \sqrt{\frac{\sum (z_i - \bar{z})^2}{n}} t_{n-1}(1 - \frac{\alpha}{2}) \tag{2}$$

   where $t_k()$ is the quantile of the T_k distribution.

Your Task:

1. For the first run in the 3-fold experiment derive a 0.95 confidence interval for the accuracy of the hypothesis obtained. Choose which approximation to use, explain your assumptions, and derive a numerical interval.

2. Derive a 0.95 confidence interval for the 3-fold estimate of the average accuracy of the hypotheses obtained by the algorithm. As before, choose which approximation to use, explain your assumptions, and derive a numerical interval.