10-602: Statistical Approaches to Learning and Discovery

Problem Set 4
Date: March 8, 2001 **Due: start of class March 15, 2001**

---

1. Suppose you have two classifiers and want to know which is better. Let p_1 be the true generalization error of classifier 1, i.e. the probability that classifier 1 will make an error on a random test point (drawn from a fixed distribution). Similarly p_2 is the true generalization error of classifier 2. On a test set of size n_1, classifier 1 happens to score e_1 errors. On a separate test set of size n_2, classifier 2 gets e_2 errors. (This more general than assuming the test sets are the same.)

   (a) Assuming the test outcomes are IID, what is your posterior distribution for p_1 and p_2? For simplicity, use the prior distribution $p(p_1, p_2) = 1$ in the box $[0, 1]^2$. Write down an integral whose value is $Pr(p_1 < p_2|e_1, n_1, e_2, n_2)$.

   (b) Implement a numerical quadrature routine for approximating the integral in part (a). For a one dimensional integral over $[0, 1]$, you could use the midpoint rule:

   $$\int_0^1 f(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} f((i - 1/2)/n) \tag{1}$$

   For a two dimensional integral, let $f(x)$ be the inner integral and apply quadrature recursively. You may want to do some analytic manipulations on the integral before applying this rule.

   Let $D = \{e_1 = 1, n_1 = 150, e_2 = 2, n_2 = 100\}$. What is $Pr(p_1 < p_2|D)$? Increase the number of quadrature points until the error is less than 10^{-2}.

   (c) Reduce this problem to homework 2 question 4 (comparison of normal distributions) by making a normal approximation to the posterior for (p_1, p_2). Do this two ways: (1) using the exact mean and variance of the posterior and (2) using a second order expansion about the mode of the posterior. What are the resulting approximations to $Pr(p_1 < p_2|D)$?

   (d) Let $o_1 = \log(\frac{p_1}{1-p_1})$ and $o_2 = \log(\frac{p_2}{1-p_2})$. The distribution of these variables is close to normal. Apply the delta method to the (first) normal approximation in part (c) to get a normal approximation for o_1 and o_2. What is the resulting approximation to $Pr(p_1 < p_2|D)$?

   (e) Let $s_1 = 2\arcsin(\sqrt{p_1})$ and $s_2 = 2\arcsin(\sqrt{p_2})$. The distribution of these variables is close to normal. Apply the delta method to the (first) normal approximation in part (c) to get a normal approximation for s_1 and s_2. What is the resulting approximation to $Pr(p_1 < p_2|D)$?

2. Let animals be distributed into four categories: $Y = (y_1, y_2, y_3, y_4)$ according to the multinomial genetic linkage model with cell probabilities: $((2+\theta)/4, (1-\theta)/4, (1-\theta)/4, \theta/4)$.

   (a) Use the Newton-Raphson algorithm to obtain the MLE for the data $Y = (14, 0, 1, 5)$. Try starting it at $\theta = 0.2, 0.3, 0.4, 0.6$, and 0.8. Does it converge for all of the starting points?

   (b) Plot the likelihood (not the log-likelihood) along with the normal approximation resulting from second-order expansion at the MLE. Scale the normal approximation so that they touch at the MLE. How appropriate is this approximation?

   (c) Now run EM for the MLE and plot the final EM bound on the same plot as part (b). It should also touch the likelihood at the MLE. How appropriate is the bound as an approximation to the likelihood?