625.664 Computational Statistics

Problem Set 8

Associated Reading: Chapter 12 from Gentle.

Complete the problems either by hand or using the computer and upload your final document to the Blackboard course site. All final submittals are to be in PDF form. Please document any code used to solve the problems and include it with your submission.

0. Read the Mathematics of a Lady Tasting Tea Handout.

1. Posted on the course Blackboard site is the data set CVdata.txt containing 2 lists, X and Y, each of length 100. When written as an ordered pair, (x, y), x is a sample observation and $y = f(x)$, the observed value of the density at x. In this problem we will use balanced half-sampling to predict the error of fitting the curve (X, Y) with a normal density, $N(\mu, 2)$. As our error metric, we will use $R(y, g) = |y - g|$.
   (a) Plot the (X, Y) and use the MLE $\hat{\mu}$ to fit $f(x)$ with a normal distribution. Clearly state the fitted function g and find the apparent error, $E_{\hat{P}_{Y|X}}$.
   (b) Partition the data set into $S_1 = \{x_1, \ldots, x_{50}\}$ and $S_2 = \{x_{51}, \ldots, x_{100}\}$ and find $E_{\tilde{P}_{Y|X}}$ (Eq. 12.5 in Gentle).
   (c) Discuss your results.

2. (a) For $r = n$, show that the jackknife variance estimate, $\widehat{V(T)_J}$ (equation 12.11 in Gentle), can be expressed as
$$\frac{n-1}{n}\sum_{j=1}^{n}(T_{(-j)} - \bar{T}_{(\bullet)})^2.$$
   (b) Again, for $r = n$, show that
$$\widehat{V(T)_J} \leq \frac{\sum_{j=1}^{n}(T_j^* - T)^2}{n(n-1)}.$$

3. The statistic
$$b_2 = \frac{\sum(y_i - \bar{y})^4}{(\sum(y_i - \bar{y})^2)^2}$$
   is sometimes used to decide whether a least squares estimator is appropriate (otherwise, a robust method may be used).
   (a) What is the jackknife estimate of the standard deviation of b_2?
   (b) Posted on the course Blackboard site is the data set Jackknife.txt containing 100 observations from a $N(0, 1)$ distribution. Use this sample to calculate b_2 and the jackknife estimate of the standard deviation for the cases $k = 1$ and $k = 5$.
   (c) Discuss the performance of the jackknife estimators found in (b). Be specific and use any techniques you feel are appropriate. (For example: You could draw several more samples from $N(0, 1)$ and use them to obtain another estimate of the standard deviation.)