Stat 8003, Homework 5

Group G: `sample( c( "David" , "Andrew", "Salam" ))`

September 29, 2014

Question 4.1. Consider a simulated dataset. Assume that the data $x_1, x_2, \cdots, x_n$ follows the following distribution:

$$x_i \sim f(x_i) = \pi_0 f_0(x_i) + \pi_1 f_1(x_i)$$

where $f_0(x_i) = 1(0 \leq x_i \leq 1)$ is the density function of the uniform and $f_1(x_i) = \beta(1 - x)^{\beta-1}$ is the density function of $Beta(1, \beta)$. The group information can be treated as a missing value and is denoted as z_i. Let $y_i = (x_i, z_i)$ be the complete data.

(a) Derive the complete likelihood function;

(b) Using the EM algorithm to derive the estimator for π_0 and β;

(c) Apply your method to the data set, estimate π_0 and β and the calculate $fdr_i = P(Z_i = 0 \mid x_i)$. (This score is called the local fdr score.)

(d) Classify x_i to the first group if $fdr_i(x_i) > 0.5$. Compare your classification with the actual group information, what is the total number of falsely classified data?

Answer:

Question 4.2. (Continued from Problem 1.) It is known that the local fdr score can be written as

$$fdr_i(x_i) = \frac{\pi_0 f_0(x_i)}{f(x_i)}$$

where $f(x_i)$ is the marginal density of x_i. Assume that $\pi = 0.7$.

(a) Estimate $f(x_i)$ by using the kernel density estimation with Gaussian kernel and Silverman's h;

(b) Estimate the local fdr score;

(c) Using the same rule as in 1(d), calculate the total number of falsely classified data;

(d) Choose the bandwidth using the maximum likelihood cross validation, repeat problem (a-c), what is the total number of falsely classified data?

(e) Which method works the best in terms of having the smallest classification error?

Answer: