CMSC 25025 / STAT 37601

Machine Learning and Large Scale Data Analysis

Assignment 7

Due: Thursday, May 30, 2013

This assignment consists of two "pencil and paper" problems and one computing problem.

1. *Leave one out* (30 points)

   As discussed in class, kernel smoother is given by

$$\widehat{Y}(x) = \frac{\sum_{j=1}^n K_h(X_j, x) Y_j}{\sum_{j=1}^n K_h(X_j, x)}$$

   for training data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Let $\widehat{Y} = (\widehat{Y}(X_1), \ldots, \widehat{Y}(X_n))^T$, and let L be the $n \times n$ matrix

$$L_{ij} = \frac{K_h(X_j, X_i)}{\sum_{j=1}^n K_h(X_i, X_j)}.$$

   The fitted values can then be written as $\widehat{Y} = LY$.

   The leave-one-out cross validation risk is defined as

$$R_{\text{LOO}}(h) = \frac{1}{n} \sum_{i=1}^n (Y_i - \widehat{Y}_{(i)})^2$$

   where

$$\widehat{Y}_{(i)} = \frac{\sum_{j \neq i} K_h(X_j, X_i) Y_j}{\sum_{j \neq i} K_h(X_j, X_i)}$$

   is the predicted value at X_i estimated by leaving the data point (X_i, Y_i) out of the sample.

   Show that the leave-one-out cross-validation risk can be written as

$$R_{\text{LOO}}(h) = \frac{1}{n} \sum_{i=1}^n \left(\frac{Y_i - \widehat{Y}_i}{1 - L_{ii}} \right)^2.$$

2. *Gaussian tails* (40 points)

   (a) If $Z \sim N(0, 1)$ show that

$$\mathbb{P}(|Z| > t) \leq \sqrt{\frac{2}{\pi}} \frac{e^{-t^2/2}}{t}$$

   To do this, first note that $\mathbb{P}(|Z| > t) = 2\mathbb{P}(Z > t)$. Next, write an expression for $\mathbb{P}(Z > t)$ and note that $x/t > 1$ whenever $x > t$.

(b) If $Z_1, \ldots, Z_n \sim N(0, \sigma^2)$ are independent, show that

$$\mathbb{P}\left(\max_{1 \le i \le n} Z_i > \sqrt{2r\sigma^2 \log n} \right) \to 0$$

as $n \to \infty$ for any $r \ge 1$.

3. *Computing is believing* (30 points)

This problem is aimed at giving a better understanding of the results above, and the thresholding technique used for the exoplanet data project.

Generate T trials of normal samples $Z_1^{(t)}, \ldots, Z_n^{(t)} \sim N(0,1)$ of size n, for $t = 1, \ldots, T$. For each trial, compute $\max_{1 \le j \le n} |Z_j^{(t)}|$. From these statistics, you can estimate the probability $\mathbb{P}(\max_{1 \le j \le n} |Z_j^{(t)}| \ge \sqrt{2r \log n})$ using the empirical probabilities over the T trials:

$$\widehat{p}_n(r) \;\equiv\; \widehat{\mathbb{P}}\Big(\max_{1 \le j \le n} |Z_j^{(t)}| \ge \sqrt{2r \log n} \Big) \;=\; \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\Big(\max_{1 \le j \le n} |Z_j^{(t)}| \ge \sqrt{2r \log n} \Big)$$

Choose a sufficiently large number of trials, such as $T = 1{,}000$, and plot $\widehat{p}_n(r)$ versus r for several values of n, such as $n = 2^k$ for $k = 5, 6, \ldots, 15$.