Homework to Week 9

Statistics: Principle, Methods and R (II)

Gao Fengnan

27th May 2017

The homework is due on Monday, 5th June 2017. Please hand in the solutions to the teaching assistant He Siyuan at the beginning of the lecture.

1. Download the `spam` data from
   `http://www.stat.cmu.edu/~larry/all-of-statistics/index.html`
   The data contain 57 covariates relating to email messages. Each email message was classified as spam ($Y = 1$) or not spam ($Y = 0$). The outcome Y is the last column in the file. The goal is to predict whether an email is spam or not.

   (a) Construct classification rules using (i) LDA, (ii) QDA, (iii) logistic regression, and (iv) a classification tree. For each, report the observed misclassification error rate and construct a 2-by-2 table of the form

	$\hat{h}(x) = 0$	$\hat{h}(x) = 1$
$Y = 0$	??	??
$Y = 1$	??	??

   (b) Use 5-fold cross-validation to estimate the prediction accuracy of LDA and logistic regression.

   (c) Sometimes it helps to reduce the number of covariates. One strategy is to compare X_i for the spam and email group. For each of the 57 covariates, test whether the mean of the covariate is the same or different between the two groups. Keep the 10 covariates with the smallest p-values. Try LDA and logistic regression using only these 10 variables.

2. Apply the k-nearest-neighbor classifier for the `iris` dataset. Choose k by cross-validation. The `iris` dataset is part of the standard R software.

3. Let $r(x) = \mathbb{P}(Y = 1 \mid X = x)$ and let $\hat{r}$ be an estimate of r. Consider

$$h(x) = \begin{cases} 1 & \text{if } \hat{r}(x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}.$$

Assume that $\hat{r}(x) \approx N(\overline{r}(x), \sigma^2(x))$ for some functions $\overline{r}(x)$ and $\sigma^2(x)$. Show that, for fixed x,

$$\mathbb{P}(Y \neq h(x)) \approx \mathbb{P}(Y \neq h^*(x)) + |2r(x) - 1| \times \left[1 - \Phi\left(\frac{\text{sign}\big(r(x) - (1/2)\big)\big(\overline{r}(x) - (1/2)\big)}{\sigma(x)}\right)\right],$$

where Φ is the standard normal CDF and h^* is the Bayes rule. Regard $\text{sign}\big((r(x)-(1/2)\big)\big(\overline{r}(x)-(1/2)\big)$ as a type of bias term. Explain the implications for the bias-variance trade-off in classification. Hint: first show that

$$\mathbb{P}(Y \neq h(x)) = |2r(x) - 1|\mathbb{P}(h(x) \neq h^*(x)) + \mathbb{P}(Y \neq H^*(x)).$$

4. Expand the following functions in the cosine basis on $[0, 1]$. For (a) and (b), find the coefficients β_j analytically. For (c) and (d), find the coefficients numerically, i.e.

$$\beta_j = \int_0^1 f(x)\varphi_j(x)dx = \frac{1}{N}\sum_{l=1} f(l/N)\varphi_j(l/N)$$

for some large integer N. Then plot the partial sum $\sum_{j=1}^n \beta_j\varphi_j(x)$ for increasing values of n.

(a) $f(x) = \sqrt{2}\cos(3\pi x)$.

(b) $f(x) = \sin(\pi x)$.

(c) $f(x) = \sum_{j=1}^{11} h_j K(x - t_j)$ where $K(t) = (1 + \text{sign}(t))/2$ with

$$(t_j) = (.1, .13, .15, .23, .25, .40, .44, .65, .76, .78, .81),$$
$$(h_j) = (4, -5, 3, -4, 5, -4.2, 2.1, 4.3, -3.1, 2.1, -4.2).$$

(d) $f(x) = \sqrt{x(1-x)}\sin\big(2.1\pi/(x + 0.05)\big)$.

5. Consider the glass fragments data from `http://www.stat.cmu.edu/~larry/all-of-statistics/=data/glass.dat` Let Y be refractive index (`RI`) and let X be aluminum content (`Al`).

(a) Do a nonparametric regression to fit the model $Y = f(x) + \varepsilon$ using the cosine basis method. The data are not on a regular grid. Ignore this when estimating the function. (But do sort the data first according to x.) Provide a function estimate, an estimate of the risk, and a confidence band.

(b) Use the wavelet method to estimate f.

6. Show that the Haar wavelets are orthonormal.

7. Consider the doppler signal:

$$f(x) = \sqrt{x(1-x)}\sin\Big(\frac{2.1\pi}{x + 0.05}\Big).$$

Let $n = 1024$, $\sigma = 0.1$, and let $(x_1, \ldots, x_n) = (1/n, \ldots, 1)$. Generate data

$$Y_i = f(x_i) + \sigma\varepsilon_i$$

where $\varepsilon_i \sim N(0, 1)$.

a) Fit the curve suing the cosine basis method. Plot the function estimate and confidence band for $J = 10, 20, \ldots, 100$.

b) Use the Haar wavelet to fit the curve.

8. (Haar density estimation.) Let $X_1, \ldots, X_n \sim f$ for some density f on $[0, 1]$. Let's consider constructing a wavelet histogram. Let φ and ψ be the Haar father and mother wavelet. Write

$$f(x) = \varphi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \beta_{jk}\psi_{jk}(x),$$

where $J \approx \log_2(n)$. Let

$$\hat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^{n} \psi_{jk}(X_i).$$

(a) Show that $\hat{\beta}_{jk}$ is an unbiased estimate of β_{jk}.

(b) Define the Haar histogram

$$\hat{f}(x) = \varphi(x) + \sum_{j=0}^{B} \sum_{k=0}^{2^j-1} \hat{\beta}_{jk}\psi_{jk}(x)$$

for $0 \leq B \leq J - 1$.

(c) Find an approximate expression for the MSE as a function of B.

(d) Generate $n = 1000$ observations from a Beta(15, 4) density. Estimate the density using the Haar histogram. Use leave-one-out cross-validation to choose B.