Homework to Week 9

Statistics: Principle, Methods and R (II)

Gao Fengnan

24 April 2017

The homework is due on Monday, 8 May 2017. Please hand in the solutions to the teaching assistant He Siyuan at the beginning of the lecture.

1. Let $X_1, \ldots, X_n \sim f$ and let $\hat{f}_n$ be the kernel density estimator using the boxcar kernel:

$$K(x) = \mathbb{1}_{(-1/2,1/2)}(x).$$

   (a) Show that

$$\mathbb{E}[\hat{f}(x)] = \frac{1}{h}\int_{x-(h/2)}^{x+(h/2)} f(y)dy$$

   and

$$\mathrm{Var}[\hat{f}(x)] = \frac{1}{nh^2}\left[\int_{x-(h/2)}^{x+(h/2} f(y)dy - \left(\int_{x-(h/2)}^{x+(h/2)} f(y)dy\right)^2\right].$$

   (b) Show that if $h \to 0$ and $nh \to \infty$ as $n \to \infty$, then $\hat{f}_n(x) \xrightarrow{\mathrm{P}} f(x)$.

2. Get the data on fragments of glass collected in forensic work from `http://www.stat.cmu.edu/~larry/all-of-statistics/=data/glass.dat`.
   Estimate the density of the refractive indexes (the variable with the label `RI`) using a histogram and use a kernel density estimator. Use cross-validation to choose the amount of smoothing. Experiment with different binwidths and bandwidths. Comment on the similarities and differences. Construct 95 percent confidence bands for your estimators. Please read page 317 of the book *All of Statistics* to learn how to construct a confidence band for the kernel density estimator.

3. Consider the data from the last question. Let Y be refractive index and let x be aluminium content (the variable with the label `Al`). Perform a nonparametric regression to fit the model $Y = f(x) + \varepsilon$. Use cross-validation to estimate the bandwidth. Construct 95 percent confidence bands for your estimate. Please read pages 321–323 of *All of Statistics* to learn how to construct confidence bands for the Nadaraya-Watson kernel estimators.

4. Consider regression data $(x_i, Y_i)_{i=1}^n$. Suppose that $0 \leq x_i \leq 1$ for all i. Define bins $(B_j)_{j=1}^m$ as in the histogram estimator. Define the estimator

$$\hat{r}_n(x) = \sum_{j=1}^{m} \mathbb{1}_{\{x \in B_j\}} \overline{Y}_j,$$

where $\overline{Y}_j$ is the average of all the Y_j's corresponding to those x_i's in B_j. Find the approximate risk of this estimator. From this expression for the risk, find the optimal bandwidth. At which rate does the risk go to zero?