Assignment : Numerical Problems

March 23, 2017

Note the following:

The use of R language is recommended for writing codes.
The use of R packages is generally discouraged (except for generating data sets).
You will get additional credits for writing your own codes.
You are free to discuss, but you will get zero if copying is established either in the writeup, or the codes.
Read and implement the instructions stated at the end of problems 1-5.
No questions will be entertained one week before the submission deadline.

Total marks: 30

---

1. Write a code for kernel density estimate (KDE) in $\mathbb{R}$ for $h = 2^a$ with $a = -4(1)5$ with the Gaussian and uniform kernels. Generate data of size $n = 100$ from uniform, exponential and Cauchy distributions for $d = 1$. Compare these estimates, and comment on what you observe.

Now, write a code for KDE in $\mathbb{R}^d$ with $d > 1$ and repeat the exercise for $d = 2, 5, 10$ and 25. Compare these estimates, and comment explicitly on what you observe. [2+2]

2. Write a code for the Nadaraya-Watson (NW) estimator in $\mathbb{R}^2$. Also, write a iterative code for the *locally linear* NW estimator. To understand how NWE compares with usual linear regression, study the following example.
Assume $Y = m(X) + \epsilon$, with $\epsilon \sim N(0, 0.1^2)$ and independent of $X \sim N(0, 0.4^2)$. Take two choices of $m(X) = 1 + X$, $0.3X + X^{23}$, e^{3+X} and $\sin X$. Simulate $n = 200$ observations (Y, X), and forget $m(X)$.
Estimate $m(X)$ with linear regression, and the two versions of NW from the data only. Compare these estimates, and give your detailed comments.

Analyze the real data set 5.4 from LW's book with the covariate as time (in milliseconds) and the response as acceleration at time of impact using NWE. State your observations.

How will you do NW regression when Y is a categorical variable? Develop this method of classification. Study the glass data set 5.3 from LW's book. [2+2+2+2]

3. Let $X_1, \ldots, X_n$ be i.i.d. from a symmetric distribution F with $F(0) = 1/2$. Consider the four statistics, namely, $T_{1n} = \overline{X}_n$, $T_{2n} = \widetilde{X}_n$, $T_{3n} = \hat{X}_n$ (the sample mode) and $T_{4n} = \widetilde{Y}_n$, where $Y_{ij} = (X_i + X_j)/2$.
The statistic T_{4n} is called the *'Hodges-Lehmann estimator'*.

Compare the efficiency of these four estimates by generating a sample of size $n = 100$ from the normal, uniform, Laplace, logistic and Cauchy distributions (all symmetric about 0) over 500 iterations. Comment on the robustness of these estimates as well. [2+3]

4. Write an algorithm to implement the exact run test. Use the codes in the R package `randtests` to execute various run tests, and study the real data in this R package. [2+3]

5. Write a code to evaluate the Kolmogorov-Smirnov and Cramer von Mises statistics in $\mathbb{R}$ for a specified null distribution F_0.

Take $F_0 \equiv N(0, 1)$. Generate data of size 100 from $N(0.05, 1)$, $N(0, 2)$ and $N(0.05, 2)$ for 500 times to compute the powers. Compare these results with the Chi-squared test. Report the average values of the power, and the respective stantard errors.
What happens if the alternative is $DE(0, 1)$, $C(0, 1)$ and $Exp(0, 1)$? [2+3]

6. Let $U_1, \ldots, U_n$ be a random sample of size n from the $U(0, 1)$ distribution. Define $G_n(t)$ to be the corresponding empirical distribution function (edf) and $Y_n(t) = G_n(t) - t$ with $t \in [0, 1]$.

Generate a random sample of size n from the $U(0, 1)$ distribution, and plot the *uniform empirical process* $\sqrt{n}\, Y_n(t)$ with $t \in [0, 1]$ for $n = 50, 100, 500$ and 1000. [3]

Submit this assignment with a two-page writeup (a single .pdf file), and codes (a single .txt file) only to the email id: **assignment.stat.iitk@gmail.com**. *Do not send to any other email id(s).*
In the subject of the email, use the format `roll number - name` *only.*

Deadline : 05.04.2016.
