Homework 3

STAT W4413: Nonparametric Statistics

DUE: Tuesday, February 23, 12:00 noon

(1) Please sign your home work with your name and UNI number.

(2) Homework must be submitted into the Statistics Homework Boxes room 904 on the 9th floor of SSW building.

(3) Homework is due Tuesday, February 23, 12:00 noon.

(4) No late homework, under any circumstances, will be accepted.

(5) At the end of semester, one of your lowest homework scores will be dropped before the final grade is calculated.

(6) Your submitted solutions should consist of (i) the hand written (or printout) of the results with all the details, (ii) printout of the relevant figures, and (iii) the printout of the source code.

1. Consider the locally polynomial regression

$$(\hat{\beta}_0(x), \hat{\beta}_1(x), \ldots, \hat{\beta}_M(x)) = \arg \min_{\beta_0, \beta_1, \ldots, \beta_M} \sum_{i=1}^{n} K\left(\frac{|x - x_i|}{h}\right) \left(y_i - \sum_{j=0}^{M} \beta_j x_i^j\right)^2.$$

Define $\hat{\beta}(x) = [\hat{\beta}_0(x), \hat{\beta}_1(x), \ldots, \hat{\beta}_\ell(x)]^T$, and $y = [y_1, \ldots, y_n]^T$. Also define the matrix X as

$$X = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^M \\ 1 & x_2 & x_2^2 & \ldots & x_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^M \end{bmatrix}$$

and also $W(x)$ as the diagonal matrix whose diagonal elements are given by $K\left(\frac{|x-x_1|}{h}\right), K\left(\frac{|x-x_2|}{h}\right), \ldots$

Prove that

$$\hat{\beta}(x) = (X^T W(x) X)^{-1} X^T W(x) y.$$

Note that as we expect the value of $\hat{\beta}$ depends on the value of x at which we estimate the function. This is in fact the major difference between locally polynomial regression and polynomial regression.

2. In this problem we would like to study the performance of the locally polynomial regression in terms of M (degree of the polynomial) and the bandwidth h. Download the dataset provided on the course website. Fit a curve to the data using the locally polynomial regression in the following two cases: (For this problem you can use locpoly function from KernSmooth package.)

   (a) Set the degree of the polynomial to 1 and fit a curve using bandwidth = 0.03, 0.1, 0.3. Compare these curves visually. Which one is a better fit? Which one do you think suffers from high variance? Which one suffers from high bias?

   (b) Now set the bandwidth to 0.3. Consider three different degree for the polynomial 1, 5, 15. Which one is a better fit? Which one suffers from high variance? Which one suffers from high bias?

3. Now consider the above example. Suppose that we set the degree of the polynomial to 1 and we would like to find the best value of the bandwidth parameter h. We do it by using 5-fold cross validation. Consider 10 equispaced values in the interval of [0, .5] for h. Use 5 fold cross validation to find the optimal value of h. Now use the optimal value of h to fit a curve to our data.

4. Image denoising is one of the most fundamental problems in image processing. In this question we would like to explore the performance of kernel regression for image denoising.

   An image can be considered as a function f whose value depends on the indices x_1, x_2. We observe a noisy observation of this function on a set of points, i.e.,

   $$y(x_1, x_2) = f(x_1, x_2) + \epsilon(x_1, x_2). \quad (x_1, x_2) \in \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$$

   For instance for the top-left corner of the image shown below $x_1 = 1, x_2 = 1$. For the top-right corner $x_1 = 1, x_2 = n$, and finally for the bottom-right corner $x_1 = n, x_2 = n$.

Similar to the nonparameteric settings our objective is to estimate f. But we only care about the value of f at locations $(x_1, x_2) \in \{1, 2, \ldots, n\} \times \{1, 2, \ldots, n\}$. Write a program that does Gaussian Kernel regression for different values of h. Assume h to be one of these number: 0.1, .5, 1, 5, 10. Estimate the image at every pixel. Which image seems visually more appealing?

Hint 1: Use "jpeg" package. It has two instructions "readJPEG" and "writeJPEG".

Hint 2: Note that you should do two-dimensional Kernel regression.

Hint 3: In case you apply the full Kernel regression your program will be slow. Set the Kernel value to zero outside a block of size 9x9 around each pixel. Do the same with a block of size 3×3. Report the differences you notice.

Hint 4: A pixel is a boundary pixel if some of the pixels in its neighborhood (9×9 or 3×3 depending on the neighborhood you consider) are out of the boundaries of the image. For those pixels take the weighted average of the pixels that exist.

5. Given a dataset $\{(X_i, Y_i)\}_{i=1}^n$ we are looking for the function that minimizes

$$\sum_{i=1}^{n} (y_i - h(x_i))^2 + \lambda \int h''(t)^2 dt. \tag{1}$$

The goal of this problem is to show that the minimizing function has the form of a natural cubic spline with knots at $X_1, X_2, \ldots, X_n$. Assume that we are interested in estimating the function on interval $[a, b]$, and that all the X_is are distinct. The first fact that we claim here without a proof is this: given a set of points (X_i, W_i), there is a **unique** natural cubic spline with knots at $X_1, X_2, \ldots, X_n$ that passes through all the (X_i, W_i) points. We can now prove the optimality of natural cubic spline by using this fact. Suppose that $\tilde{g}$ is the function that achieves the minimum of (1). Define $Z_i = \tilde{g}(X_i)$. Let g be a natural cubic spline with knots at $X_1, X_2, \ldots, X_n$ that passes through $\{(X_i, Z_i)\}_{i=1}^n$. We want to prove that if we plug g in (1) we get a smaller result.

(a) Prove that $\sum_{i=1}^n (y_i - g(x_i))^2 = \sum_{i=1}^n (y_i - \tilde{g}(x_i))^2$.

(b) Let $h(x) = \tilde{g}(x) - g(x)$. Use integration by parts to prove that

$$\int_a^b g''(x)h''(x)dx = -\sum_{j=1}^{N-1} g'''(x_j^+)(h(x_{j+1}) - h(x_j)).$$

Hint 1: $g''(a) = g''(b) = 0$. Why?
Hint 2: $g(x_j^+)$ is the limit of $g(x)$ as x goes to x_j from the right side.

(c) Calculate the value of $h(x_j)$. Use this value and part (a) to calculate $\int_a^b g''(x)h''(x)dx$.

(d) According to Cauchy-Scwartz inequality we have

$$\left|\int_a^b g_1(x)g_2(x)\right|^2 dx \leq \int_a^b g_1^2(x)dx \int_a^b g_2^2(x)dx.$$

Also, equality holds if and only if $g_1(x) = \alpha g_2(x)$ for some constant α. Use Cauchy Schwartz inequality and part (c) to prove that

$$\int_a^b (\tilde{g}''(x))^2 dx \geq \int_a^b g''(x)^2 dx.$$

Also, prove that equality holds if and only of $\tilde{g}(x) = g(x)$.

Note: Cauchy-Schwartz inequality is the inequality that you have used in basic probability courses to prove that the correlation coefficient is always between -1 and 1.

(e) Combine all the results you have proved in (a), (b), (c), and (d) and prove that the function that minimizes (1) has the form of a natural cubic spline.