ECE 411 Homework 6 (Fall 2023)
Instructor: Dr. Chau-Wai Wong
TA in Charge: Mr. Prasun Datta
Material Covered: Statistical Learning Basics

Problem 1 (20 points) (20 points) [Curse of Dimensionality] Read the first paragraph of the problem statement of *ESLII-2.4*. Note that we may also write $\mathbf{X} = [X_1, X_2, \ldots, X_p]^T$, where $X_k \sim \mathcal{N}(0, 1)$ for $k = 1, \ldots, p$. Use a programming language of your choice. To get started, set $p = 10$. Note that in this problem, all vectors are column vectors.

a) Write a computer program to randomly draw/generate $N = 100$ vectors from the template random vector $\mathbf{X}$, namely, $\{\mathbf{x}^{(i)},\ i = 1, \ldots, N\}$. Note that each vector should contain p normally distributed random numbers. Plot all vectors as points in a 3-D space consisting of the first, second, the last coordinates.

b) Calculate the coordinate value of each point after being projected onto a fixed direction specified by $\mathbf{a} = \mathbf{x}_0/\|\mathbf{x}_0\|$, namely, $z^{(i)} = \mathbf{a}^T\mathbf{x}^{(i)}$. Here, $\mathbf{x}_0$ is an arbitrary nonzero vector of length p, "T" is the transpose operation, and $z^{(i)} \in \mathbb{R}$. What are the sample mean and sample variance of the projected coordinates $\{z^{(i)},\ i = 1, \ldots, N\}$?

c) Repeat a) and b) for $p \in [1, 80]$. You may want to use a `for` loop to achieve this. Optionally, put your code for parts a) and b) into a function to make your code easier to read. Plot the sample variance of the projected coordinates as a function of p.

d) Calculate the squared distance of each point to the origin, namely, $d_i^2 = \|\mathbf{x}^{(i)}\|^2$. What is the sample mean of $\{d_i^2,\ i = 1, \ldots, N\}$? Plot the sample mean of the squared distance as a function of p in the same plot of c). Limit the range of y-axis between 0 and 80. For $p = 5$, inspect the values of any five d_i^2's. Do the results in b) and c) match with the conclusion drawn in the third paragraph of *ESLII-2.4*?

e) Use the formulas from Problem 3b of HW5, prove that $\mathrm{Var}(Z) = 1$ where $Z = \mathbf{a}^T\mathbf{X}$, and $\mathbb{E}[D^2] = p$ where $D = \|\mathbf{X}\|$. Are the theoretical results in this part consistent with the simulated results obtained in c) and d)?

Problem 2 (20 points) [Alternative Neighbor Averaging Method for Simulated Data]

a) Given a regression function $f(x) = x^2 + 2x + 1$ and a linear model $Y = f(X) + e$, where $e \sim N(0, 1)$ and $X \sim \mathrm{Uniform}(-1, 1)$, generate 50 pairs of (x_i, y_i) and graph them using black circles. Also, plot the regression function using a black solid curve.

b) We use a method similar to the nearest neighbor averaging to estimate the regression function. We use a neighborhood of fixed radius $\delta = 0.1$. The estimated regression function takes the following form:

$$\hat{f}(x) = \frac{1}{|I(x)|} \sum_{i \in I(x)} y_i, \quad I(x) = \{i : |x - x_i| \leq \delta\}, \tag{1}$$

where $I(x)$ is the set of indices of x_i such that they are within δ in terms of distance from x, and $|I(x)|$ is the number of elements of set $I(x)$. For example, when $x = 0.9$ and $\delta = 0.1$, you first need to find all points that are within the range of [0.8, 1.0] in the x-direction, and then take the average of their values in the y-direction to obtain $\hat{f}(0.9)$. You may want to calculate $\hat{f}(\cdot)$ for all $x \in [-0.9, 0.9]$ with a step size of 0.01. If there is not a single point within the current neighborhood, use the $\hat{f}$ from the previous step as that for the current step. Draw the estimated regression function using a red solid curve in the same plot of a).

c) (Bonus, 5 points) Vary the neighborhood radius δ, how does the shape of the estimated regression function change?

Problem 3 (20 points) [Linear Regression with R] Complete *ISLR-3.6.1–3, 3.6.7, 3.7.8.*

For Python users, please download *Boston.csv data* and follow the text book's instructions while referring to *the "equivalence" Python codes of ISLR-3.6.1–3, 3.6.7* and *of ISLR-3.7.8*, where you may find the comments useful.

Problem 4 (Bonus, 20 points) [Interpretation of Confidence Interval]

a) Given a regression function $f(x) = 3x + 1$ and a linear model $Y = f(X) + e$, where $e \sim N(0, 1)$ and $X \sim \text{Uniform}(-1, 1)$, generate 50 pairs of (x_i, y_i).

b) Using the equations in the lecture slides, calculate all the estimates, namely, $\hat{\beta}_0$ and $\hat{\beta}_1$, and their standard errors. Note that when calculating standard errors, use the estimated value $\text{RSS}/(n-2)$ to replace the theoretical quantity σ^2.

c) Calculate the confidence interval for β_1. Is 3 included in the interval?

d) Repeat (a)–(c) 1000 times. What is the chance that 3 is included in a calculated confidence interval?

e) Now, can you explain what is a confidence interval?

Problem 5 (Bonus, 20 points) [Hypothesis Test] Download the `advertising` dataset from ISLR's website. Using formulas from the lecture slides, manually fit a linear model using `sales` against `TV` and with intercept. Re-calculate all entries of tables on slides 11 and 13, except the F-statistics. When calculating the p-values, use a standard normal distribution instead of the t distribution. Are the re-calculated values consistent with those in the tables?