ECE 492-45 Homework 6 (Fall 2021)
Instructor: Dr. Chau-Wai Wong
Material Covered: Statistical Learning Basics

Problem 1 (20 points) [Optimality of Mean Operators]

a) We are given two variables X and Y that are not independent. Hence, we may use one to estimate the other. Find the best deterministic function $g(\cdot)$ such that it minimizes the expected squared error between Y and $g(X)$ conditioned on $X = x$. You may find a change of variable using θ in the place of $g(x)$ helpful. Pay attention to write clearly the upper case X and the lower case x in your submission.

b) *Arithmetic average*, or the *sample mean* in a statistical context, is commonly used in everyday life for making quantitative description. We examine a statistical interpretation for the arithmetic average below. A person weighs μ lb. He tried multiple scales in a supermarket and recorded the reading from each scale, denoted by Y_i for the ith scale. We may create a linear model as follows to relate the true weight μ and the measurement Y_i:

$$Y_i = \mu + e_i, \quad i = 1, \ldots, N,$$

where e_i is the measurement error of the ith scale. Use the mean-square criterion $J(\mu) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2$ to find the closed-form expression for the best estimator for μ. The expression should contain $\{Y_i\}_{i=1}^{N}$ only, and should not contain such symbols as μ or e_i as they were not available when readings were recorded. Does the expression make intuitive sense?

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 $X \sim \text{Uniform}(-1, 1)$, 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 stepsize 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.

(You are only given 3 required problems. The rest of time should be devoted to the project proposal.)

Problem 4 (20 points, bonus) [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) Use the equations in the lecture slides, calculate the all 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 a calculated confidence interval?

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

Problem 5 (20 points, bonus) [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, using a standard normal distribution in lieu of the t distribution. Are the re-calculated values consistent with those in the tables?