Stat/CS 5525, Fall 2023
Take-home Final Exam

(Due on December 12th, 2023)

Instructions:

i. *You MUST finish the exam INDEPENDENTLY without discussing with others.*

ii. *Show all your work to justify your answers.*

iii. *For the computational problems, try to make inferences on the results when answering the questions. Do not only show the computer outputs.*

iv. *Submit your answer sheet and programming code to Canvas course website by **December 12, 2023**.*

Problem 1.
Let $(Z_1, Y_1), \dots, (Z_n, Y_n)$ be generated as follows:

$$Z_i \sim \text{Bernoulli}(p)$$
$$Y_i \sim \begin{cases} N(0,1) & \text{if } Z_i = 0 \\ N(5,1) & \text{if } Z_i = 1 \end{cases}$$

(a) Generate such a dataset of (Z, Y) with size $n = 100$ and $p = 0.3$.

(b) Conduct a numerical method for parameter estimation of p based on data in (a).

Problem 2.
Use the LA ozone dataset. Divide the dataset into two groups at random. One group, which we call the training data, containing 2/3 of the observations and one group, which we call the test data, with 1/3 of the observations. In the following you are asked to regress the *cube root* of the ozone concentration on the other variables. You should *only* use the training data for the estimation.

(a) Use the lasso method to analysis the data and return the following two plots: (1) the path of the estimated coefficients with respect to λ varying from small to large. (2) the path of Training error and Test error with respect to λ varying from small to large.

b) Use the ridge regression and plot the Training error and Test error with respect to λ varying from small to large.

Problem 3.
There are two data sets, “Pr4_training.txt” as training data set, “Pr4_test.txt” as test data set. It contains two input variables X_1, X_2 and one output variable Y. Use local constant and local linear

regression to fit the training data, and make predictions at the test data, then compute the root mean square prediction error:

$$RMSPE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}(x_i))^2}.$$

Compare the two methods and make inference. (*Hints:* can the *"np"* package of R.)

Problem 4.

The traing and test data sets can be found from 'geno_train.txt' and 'geno_test.txt'. Each contain 16 columns of data from different individuals, with the first 15 being the genetic fingerprint (the count of the number of repeats for certain so-called tandem repeats in the genome) and the last being the population variable. The purpose is to predict the population from the genetic fingerprint. We refer below to the repeat counts as the count data (the x variables) and the population as the group (the y variable).

(a) Using the training data set, for each of the input variable, estimate the density for different population class, and then plot the 3 estimated density pdf's for each variable in the same plot. Use different color for different pdf's. You will have 15 plots and try to arrange all the 15 plots in a 3×5 matrix.

(b) Compare two different classification methods LDA and SVM according to the misclassification rate on the test data set.