ECE 492-45 Introduction to Machine Learning
2019 Fall Exam 1
Instructor: Dr. Chau-Wai Wong

This is a closed-book exam. You may use a scientific calculator with cleared memory, but not a smart phone or computer. You should answer ***all four*** problems.

Problem 1 (30 pts) An ECE student named Tom plans to test the fuel economy of his car in terms of how many gallons is needed for driving one mile. He will do four 4 test drives of x_i miles each, $i = 1, \cdots, 4$, and will measure the corresponding gas consumption Y_i gallons, $i = 1, \cdots, 4$ using a meter connected to his car's microcontroller. Denote the ground-truth fuel economy as k gallon/mile.

(a) Tom believes that the readings of the gas consumption Y_i are inaccurate but unbiased, so he set up a linear model $Y_i = kx_i + e_i$, $i = 1, \ldots, 4$, where e_i are measurement noise with zero-mean and variance σ^2. Express this model in the matrix-vector form. Explicitly define $\mathbf{y}$, $\mathbf{X}$, β, and $\mathbf{e}$.

(b) Use the normal equation $\mathbf{X}^T\mathbf{X}\hat{\beta} = \mathbf{X}^T\mathbf{y}$ to directly obtain the analytic form of the least-squares estimator $\hat{k}$ for the fuel economy, and simplify $\hat{\beta}$ up to a point that it cannot be further simplified. Show that $\hat{k}$ is unbiased, and derive its variance. (Hint: x_i's are constants whereas Y_i's are random variables.)

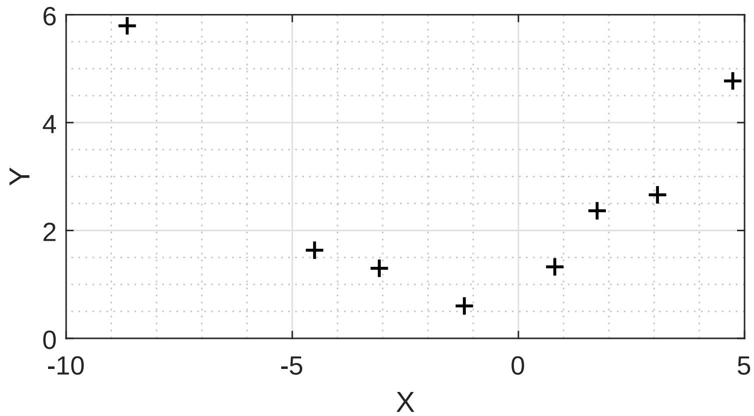
(c) Tom's brother proposed another way to estimate the fuel economy: $\tilde{k} = \left(\sum_{i=1}^{4} Y_i\right) / \left(\sum_{i=1}^{4} x_i\right)$. Show that $\tilde{k}$ is also unbiased, and derive its variance.

(d) Tom plans to drive 1, 2, 2, and 3 miles for each test drive, respectively. Compare numerically the variance of the two estimators. Is the least-squares estimator better than the one proposed by Tom's brother?

Problem 2 (20 pts) This problem investigates nearest neighbor regression and classification.

(a) Draw an estimated regression function as horizontal line segments using 1-NN rule for the data shown in the figure on page 2. Note that X is the predictor and Y is the response. Annotate the locations of the discontinuities of the estimated regression function using vertical dotted lines.

(b) In this part, class 1 should be denoted by "$\circ$", and class 2 should be denoted by "$\times$". You are given a set of training data, in which points $(1, 1)$, $(1, 2)$, and $(2, 2)$ belong to class 1, whereas points $(2, 0)$ and $(1, -1)$ belong to class 2. You are also given two test points: $(0, -1)$ and $(1.5, 0)$. 1-NN will be used for predicting their classes.



(i) For every test point, draw a table showing their distance to every point in the training data. Use the tables to determine the classes for the test points.

(ii) Draw data points on a 2D plane using "○" and "×". Label your x- and y-axes. Show tick and tick labels for each axis. Draw the 1-NN decision boundary that consists of 3 linear pieces for the training data.

(iii) What are the exact coordinates for the 2 turning points on the 1-NN decision boundary? Show your calculation steps formally (equations) or informally (drawings with numbers) to get full points.

Problem 3 (30 pts) Some R commands on the `Boston` dataset and excerpts from the R outputs are shown as follows:

```
> lm_fit0 = lm(medv ~ 1)
> summary(lm_fit0)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.5328     0.4089   55.11   <2e-16 ***

> lm_fit1 = lm(medv ~ lstat)
> summary(lm_fit1)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 34.55384    0.56263   61.41   <2e-16 ***
lstat       -0.95005    0.03873  -24.53   <2e-16 ***

> lm_fit2 = lm(medv ~ lstat + I(lstat^2))
> summary(lm_fit2)
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 42.862007   0.872084   49.15   <2e-16 ***
lstat       -2.332821   0.123803  -18.84   <2e-16 ***
I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***

> anova(lm_fit1, lm_fit2)
```

```
Analysis of Variance Table
Model 1: medv ~ lstat
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    504 19472
2    503 15347  1    4125.1 135.2 < 2.2e-16 ***

> anova(lm_fit0, lm_fit2)
Analysis of Variance Table
Model 1: medv ~ 1
Model 2: medv ~ lstat + I(lstat^2)
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    505 42716
2    503 15347  2     27369 448.51 < 2.2e-16 ***

> confint(lm_fit1)
                2.5 %     97.5 %
(Intercept) 33.448457 35.6592247
lstat       -1.026148 -0.8739505
```

(a) What is the difference between `lm_fit1` and `lm_fit2`? What is the difference between `lm_fit1` and `lm_fit0`?

(b) What is the hypothesis testing conducted using `anova(lm_fit1, lm_fit2)`? What are H_0 and H_A, respectively? Was this test result summarized in p-value also shown in the result for `lm_fit0`, `lm_fit1`, or `lm_fit2`? If yes, identify the particular line that shows the equivalent result. If no, please explain.

(c) Can the result for the hypothesis testing conducted using `anova(lm_fit0, lm_fit2)` found anywhere else? If yes, identify the particular line that shows the equivalent result. If no, please explain.

(d) Denote the true coefficient for `lstat` as β_1. Determine True or False for the following statements:

1. There is a 95% chance that the estimate for β_1, i.e., $\frac{(-1.03)+(-0.87)}{2} = -0.95$, is correct.
2. There is 95% confidence that $[-1.03, -0.87]$ contains β_1.
3. If the same procedure is repeatedly carried out and each time with the data drawn from the same population/distribution, there is 95% probability that $[-1.03, -0.87]$ contains β_1.

(e) Show in steps that the statement "β_1 is in $[\hat{\beta}_1 - 2\,\mathrm{se}(\hat{\beta}_1), \hat{\beta}_1 + 2\,\mathrm{se}(\hat{\beta}_1)]$ with 95% chance" can be written as

$$\mathbb{P}\left[\left|\frac{\hat{\beta}_1 - \beta_1}{\mathrm{se}(\hat{\beta}_1)}\right| \leq 2\right] = 0.95.$$

Problem 4 (20 pts) This problem investigates the curse of dimensionality.

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, X. We assume that X is uniformly distributed on $[0, 1]$. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of X closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.3$, we will use observations in the range $[0.25, 0.35]$. On average, what fraction of the available observations will we use to make the prediction?

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, X_1 and X_2. We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict a test observation's response using only observations that are within 10% of the range of X_1 and within 10% of the range of X_2 closest to that test observation. On average, what fraction of the available observations will we use to make the prediction?

(c) Generalize the cases in (a) and (b) to $p = 100$. What fraction of the available observations will we use to make the prediction?

(d) Using your answers to (a)–(c), comment on a drawback of k-NN when p is large.

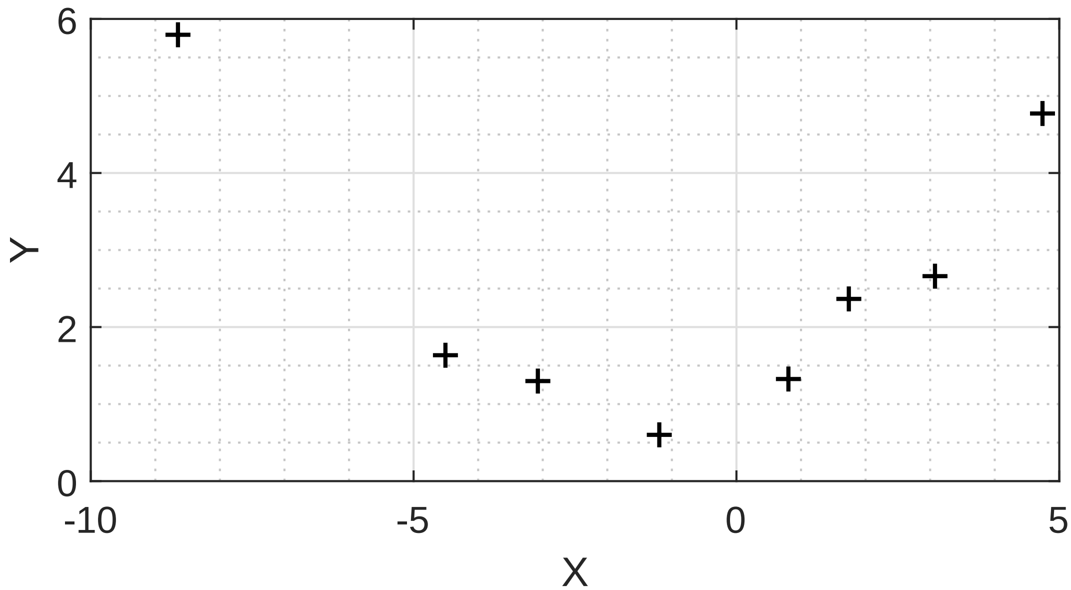
(e) Now suppose that we wish to make a prediction for a test observation by creating a p-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube?

Name:
Student ID:

Answer to Problem 1:

Name:

Answer to Problem 2(a):



Answer to Problem 2(b):

Name:

Answer to Problem 3:

Name:

Answer to Problem 4: