ECE 492-45 Introduction to Machine Learning
2021 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 (25 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 test drives of x_i miles each, $i = 1, \ldots, 4$, and will measure the corresponding gas consumption Y_i gallons, $i = 1, \ldots, 4$ 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 $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\mathbf{e}$.

(b) Use the normal equation $\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{y}$ to directly obtain the analytic form of the least-squares estimator $\hat{k}$ for the fuel economy, and simplify $\hat{\boldsymbol{\beta}}$ 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 friend proposed another way to estimate the fuel economy: $\tilde{k} = \frac{1}{4}\sum_{i=1}^{4} \frac{Y_i}{x_i}$. Examine whether $\tilde{k}$ is unbiased. Derive the variance of $\tilde{k}$.

(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 friend?

Problem 2 (25 pts) This problem investigates nearest-neighbor regression. A set of 6 training data points is drawn in the figure on page 2. Using the k-nearest-neighbor regression rule, an estimated regression function can be written as follows:

$$\hat{y}^{(k)} = \hat{f}^{(k)}(x) = \frac{1}{k}\sum_{i \in I(x)} y_i, \quad I(x) = \{i : \text{the indices of } k \text{ smallest } |x_i - x|\}. \tag{1}$$

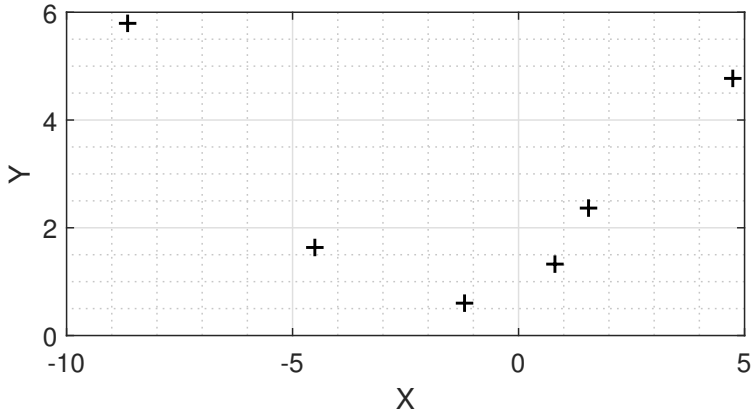
Note that this is a regression problem with only one predictor and the graphical representation of the estimated regression function on the xy-plane will be a collection of horizontal line segments.

(a) Draw the regression function $\hat{f}^{(k)}(x)$ for $x \in [-10, 5]$ when only $k = 1$ nearest neighbor is contributing to the regression.

(b) Draw the regression function $\hat{f}^{(k)}(x)$ for $x \in [-10, 5]$ when two $k = 2$ nearest neighbors are contributing to the regression.

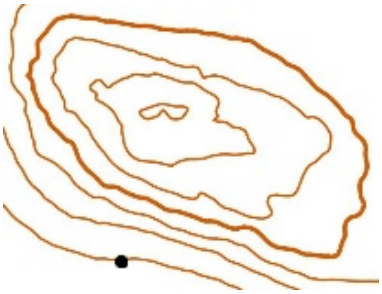
(c) Comment on how the shape of regression function will change as the number of contributing neighbors increases.

To get full points, you must annotate the locations of the discontinuities of each estimated regression function using vertical dotted lines.



Problem 3 (25 pts)

(a) A set of level curves is shown as follows. Use the dot as the starting point, draw a trajectory of gradient descent steps. Annotate each descent step using a line segment with an arrow at the end. Explicitly draw a tangent line at each step, which can assist you to determine the negative gradient direction. You may vary the descent step size.



(b) Explain how Transformer neural networks contextualize/"pay attention to" the embeddings of the sentence "walk by river bank" based on the initial embeddings that are not

aware of the context. Keep your explanation simple by ignoring the projections to the "key," "value," and "query" semantic subspaces.

Problem 4 (25 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?

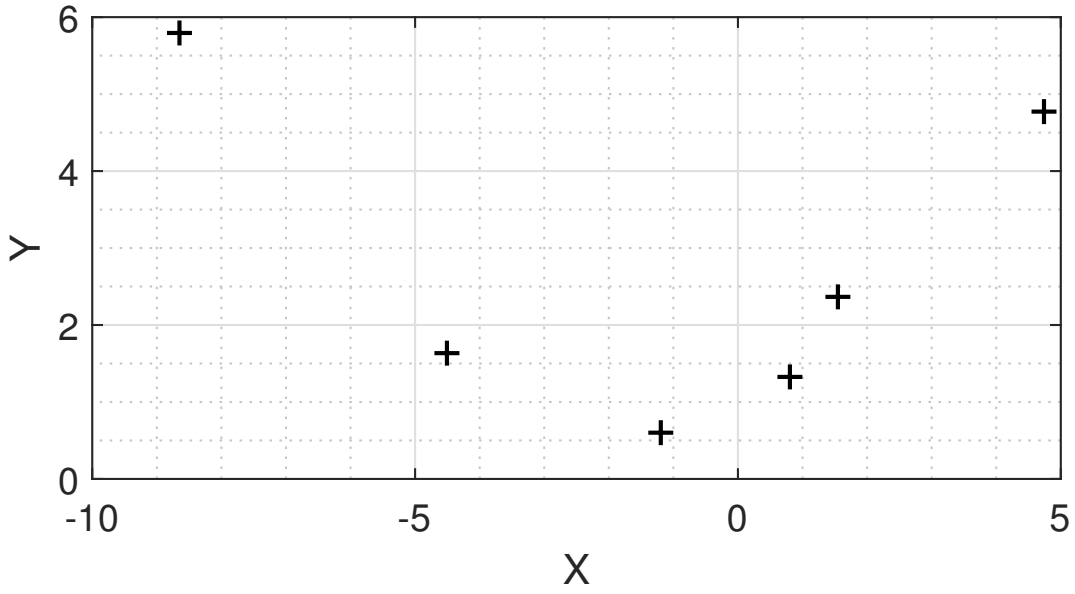
[This Page Intentionally Left Blank]

Name:
Student ID:
Answer to Problem 1:

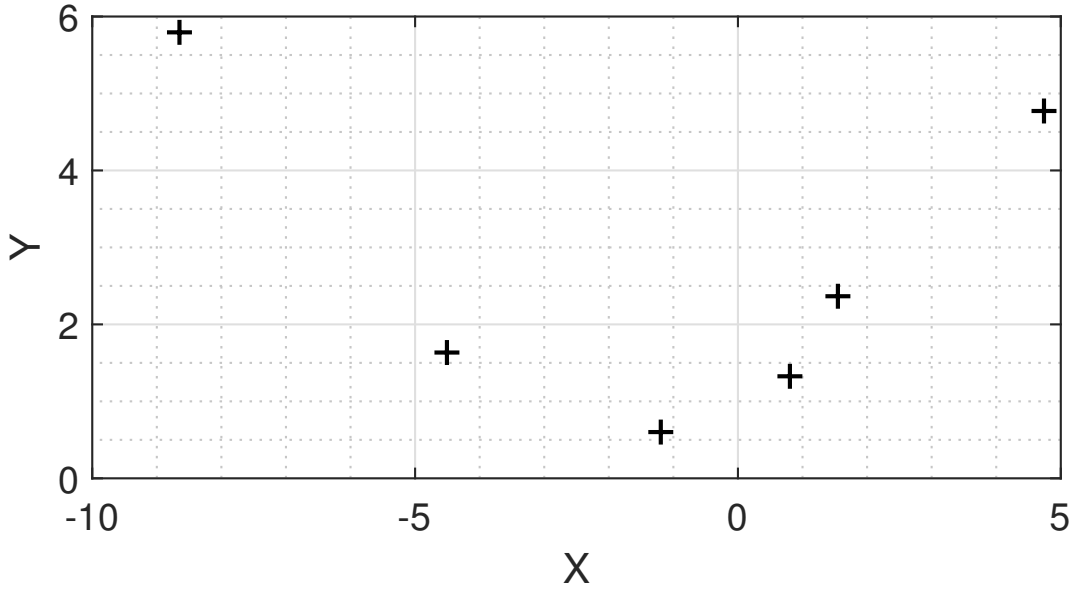
Answer to Problem 1 (cont'd):

Name:

Answer to Problem 2(a):



Answer to Problem 2(b):



Answer to Problem 2(c):

Answer to Problem 2 (cont'd):

Name:

Answer to Problem 3(a):

Answer to Problem 3(b):

Answer to Problem 3 (cont'd):

Name:

Answer to Problem 4:

Answer to Problem 4 (cont'd):