# ECE 411 Introduction to Machine Learning
# 2022 Fall Exam 2
# 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. Two-sided letter-sized handwritten cheatsheet is allowed. You should answer ***all four*** problems.

**Problem 1** (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 $[-1, 1]$. 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 $X = 0.2$, we will use observations in the range $[0.15, 0.25]$. 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 $[-1, 1] \times [-1, 1]$. We wish to predict a test observation's response using only observations that are within 20% of the range of $X_1$ and within 20% 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, 20% of the training observations. For $p = 1, 2$, and 100, what is the length of each side of the hypercube?

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

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

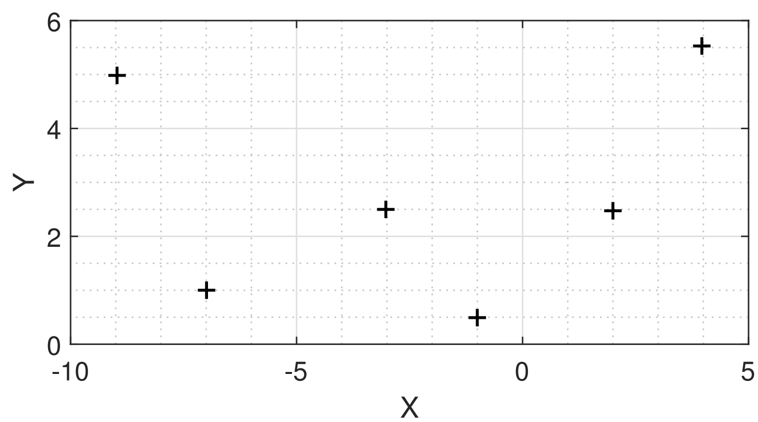
Note that this is a regression problem with only one predictor and the graphical representation of the estimated regression function on the $xy$-plane will be a collection of horizontal line segments.

**(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)** You are given random variables $\{X_i\}_{i=1}^n$ and nonzero constants $\{a_i\}_{i=1}^n$. Prove the following equality:

$$\mathrm{Var}(a_1 X_1 + a_2 X_2) = a_1^2 \mathrm{Var}(X_1) + a_2^2 \mathrm{Var}(X_2) + 2a_1 a_2 \,\mathrm{Cov}(X_1, X_2). \tag{2}$$

There are multiple proving routes. You may find some of the following expressions useful: i) the definitions $\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2]$ and $\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$; and ii) the shortcut formulas $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}X)^2$ and $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$.

**(b)** 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, where $Y_i$ has a variance of $\sigma^2$ for all $i$. Denote the ground-truth fuel economy as $k$ gallon/mile. Tom and his friends used intuition and applied some principled approaches and came up with the following candidate estimators for the fuel economy $k$:

$$\hat{k} = \left( \sum_{i=1}^{4} x_i Y_i \right) \Big/ \left( \sum_{i=1}^{4} x_i^2 \right), \tag{3a}$$

$$\tilde{k} = \frac{1}{4} \sum_{i=1}^{4} Y_i / x_i, \tag{3b}$$

$$\check{k} = \left( \sum_{i=1}^{4} Y_i \right) \Big/ \left( \sum_{i=1}^{4} x_i \right). \tag{3c}$$

(i) Derive the analytic forms of the variance of estimators $\hat{k}$, $\tilde{k}$, and $\check{k}$.

(ii) Let $(x_1, x_2, x_3, x_4) = (2, 1, 2, 3)$. Numerically calculate the variance of all estimators.

(iii) In Exam 1, you have proved that all three estimators are unbiased estimators. Based on the results in (ii), argue which one may be the best for Tom and his friends to use. Please give justification in your own words.

**Problem 4** (25 pts) Response $Y_i \sim \mathrm{B}(n, p_i)$ is a binomial random variable in which $n$ is known. The (conditional) PDF is shown as follows:

$$\mathbb{P}[Y_i = k | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = \binom{n}{k} p_i^k (1 - p_i)^{n-k}, \quad k \in \{0, 1, \ldots, n\}. \tag{4}$$

**(a)** Explain why the linear regression may not the best fit to find the relation between $Y_i$ and a set of predictors $X_{i,1}, \ldots, X_{i,q}$.

**(b)** One proposes to link the conditional mean $\mu_i$ and the predictors $\underset{\sim}{x}_i$ using a generalized linear model shown as follows:

$$g(\mu_i) = \underset{\sim}{\beta}^T \underset{\sim}{x}_i \tag{5}$$

where $g(u) = \log(\frac{u}{n-u})$ and $\mu_i = \mathbb{E}[Y_i | \underset{\sim}{X}_i = \underset{\sim}{x}_i] = np_i$. From the variable transformation viewpoint, show that $g(\cdot)$ matches the ranges for the two sides of Eq. (5).

**(c)** Rewrite the PDF into an exponential family form shown as follows:

$$f_Y(y; \theta) = \exp\left( \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right), \tag{6}$$

where $\theta$ is the natural parameter. Show that $g(\cdot)$ in (b) is the canonical link function when taking $\mu_i$ as the input.