STAT 4255
Spring 2021
Take-home Quiz

Name (Print): _____

Student ID: _____

Instructions:

- Enter all requested information on the top of this page, and put your initials on the top of every subsequent page, in case the pages become separated.
- Be sure to convert your answered quiz into PDF format before submission to HuskyCT.
- This exam contains 4 pages (including this cover page). Check to see if any pages are missing.
- The exam is open book. You may use your books or notes.

Problem I. True or False Questions. This problem is composed of 8 true-or-false statements. You only need to classify these as either `true` or `false`. No explanation is required. Remember to *mark* your answers (T or F) in the bottom table!

1. Regression problems have only quantitative inputs.
2. Regression models have quantitative outputs.
3. The training error usually decreases as we increase the model flexibility.
4. The Bayes classifier can not be implemented in practice, but if it could it would always attain zero test error.
5. Logistic regression is equivalent to linear regression if all outputs in the training data take the value 0 or 1.
6. The model bias of k-NN typically increases as k increases.
7. The model $Y = \log(x_1^{\beta_1} x_2^{\beta_2} e^{\beta_0}) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ can be estimated by linear regression, where β_0, β_1 and β_2 are the unknown coefficients and $x_1, x_2 > 0$.

   Note: e stands for the Euler number and log *means the natural logarithm.*
8. Consider $Y = 1$ to be the positive class and $Y = 0$ to be the negative class. Recall that the logistic regression classifier is

$$\widehat{C}(X) = \begin{cases} 1 & \text{if } \widehat{p}(X) > \texttt{threshold}, \\ 0 & \text{otherwise}, \end{cases}$$

   where $\widehat{p}(X)$ is the predicted class probability. The false positive rate for this classifier will always be at least as large for $\texttt{threshold} = 0.25$ as for $\texttt{threshold} = 0.5$.

Question	1	2	3	4	5	6	7	8
Answer								

Problem II. Suppose you have regression data generated by a polynomial of degree 3. Characterize the bias-variance of the estimates of the following models on the data with respect to the true model by circling the appropriate entry.

	Bias	Variance
Linear regression	low / high	low / high
Polynomial regression with degree 3	low / high	low / high
Polynomial regression with degree 10	low / high	low / high

Problem III. Explain briefly ($\sim \frac{1}{2}$ page) the meaning of the bias-variance tradeoff, i.e., what do we mean by model bias and model variance, and why is there a trade-off between the two?

Problem IV. A friend of yours is faced with a regression problem with two possible inputs, X_1 and X_2. She considers two linear regression models:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon, \tag{M1}$$
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \tag{M2}$$

Both models are fitted to a training data set $\mathcal{T} = \{(x_i, y_i)\}_{i=1}^N$ using least-squares, resulting in the two prediction models

$$\hat{Y} = 12.9 + 3.2 X_1, \tag{P1}$$
$$\hat{Y} = 11.6 - 1.4 X_1 + 1.72 X_2, \tag{P2}$$

respectively. Your friend is puzzled by these results and comes to you for advice. She says:

> "In model (P1) a unit increase in X_1 results in an *increase* of the predicted output by 3.2 units, i.e. it is clear that Y is positively correlated with X_1. However, in model (P2) a unit increase in X_1 instead results in a *decrease* of 1.4 units in the predicted output, i.e. now X_1 appears to be negatively correlated with Y!"

Give a plausible explanation to your friend's dilemma.