STATS 503, Homework 2

Please use a Jupyter notebook to write up your solutions. Submit your work through Canvas by uploading an html file that contains the html export of your notebook file. The html file should include all code (no cells hidden). Use markdown formatting to make clear which part of your notebook corresponds to each problem.

1. The table below provides a dataset containing 6 observations, 3 predictors, and 1 qualitative response variable. We will refer to this dataset as $\mathcal{D}$.

Obs.	X_1	X_2	X_3	Y
#1	0	3	0	Red
#2	2	0	0	Red
#3	0	1	3	Red
#4	0	1	2	Green
#5	-1	0	1	Green
#6	1	1	2	Red

   Consider a new test point $x^{(\text{te})}$ with $x_1^{(\text{te})} = x_2^{(\text{te})} = x_3^{(\text{te})} = 0$, that is, $x^{(\text{te})} = (0, 0, 0)^\top$. Now consider using the K-nearest neighbors (KNN) majority decision rule with $\mathcal{D}$ to make predictions about Y given $x^{(\text{te})}$.

   For any number of neighbors $K \in \mathbb{N}$ and any input values x, let $\hat{y}^{(K)}(x; \mathcal{D})$ denote the KNN prediction about the test point x based on the dataset $\mathcal{D}$.

   (a) **[3 POINTS]** Compute the squared Euclidean distance between $x^{(\text{te})}$ and each observation in $\mathcal{D}$. Report the 6 squared distances. Do not use a computer to solve this problem. (You may be asked to solve this sort of problems in an exam setting where you do not have access to a computer.)

   (b) **[2 POINTS]** What is $\hat{y}^{(3)}(x^{(\text{te})}; \mathcal{D})$? Explain your reasoning.

   (c) **[2 POINTS]** What is $\hat{y}^{(1)}(x^{(\text{te})}; \mathcal{D})$? Explain your reasoning.

   (d) **[4 POINTS]** Consider two new datasets $\mathcal{D}_1$ and $\mathcal{D}_2$, each with 6 observations, sampled by the same data-generating process that $\mathcal{D}$ was.

   True or false: for a typical data-generating process,

$$\Pr\left(\hat{y}^{(3)}(x^{(\text{te})}; \mathcal{D}_1) = \hat{y}^{(3)}(x^{(\text{te})}; \mathcal{D}_2)\right) > \Pr\left(\hat{y}^{(1)}(x^{(\text{te})}; \mathcal{D}_1) = \hat{y}^{(1)}(x^{(\text{te})}; \mathcal{D}_2)\right).$$

   Explain your reasoning.

   Note: I say "typical" above because technically either probability could be higher depending on the data-generating distribution. However, there is an answer that is correct for most real-world problems. That is the answer to give.

2. **[2 POINTS]** You have a dataset. You are considering a family of different methods you could apply to this dataset in order to estimate a regression function. The methods range from very inflexible (only capable of representing a small class of true regression functions) to very flexible (capable of fitting a very large class of true regression functions).

(a) True or false: typically, predictive error on training data will be lower with more flexible methods.

(b) Describe how predictive error on testing data changes as you use more and more flexible methods (in typical cases).

3. Assume that a predictor $X \in \mathbb{R}$ and a response $Y \in \mathbb{R}$ are governed by the data-generating process

$$X \sim \text{Uniform}[0, 1]$$
$$[Y|X = x] \sim \text{Uniform}[x + \cos(2\pi x) - .1, x + \cos(2\pi x) + .1]$$

Let $f(x) = \mathbb{E}[Y|X = x]$ denote the true regression function.

Let $\hat{f}(x; \mathcal{D})$ denote the ordinary least squares estimate of $f(x)$, based on a dataset $\mathcal{D}$. That is, $\hat{f}(x; \mathcal{D})$ can be computed by fitting ordinary least squares to the dataset $\mathcal{D}$ and then using the fitted model to make a prediction for the input x.

In this problem, we will assess how well $\hat{f}$, fitted to a dataset of 100 samples drawn from the data-generating process above, estimates f.

(a) **[1 POINTS]** Calculate $f(0.5)$. *Hint: the mean of the* Uniform$[a, b]$ *distribution is given by* $(b + a)/2$.

(b) **[1 POINTS]** Calculate $\text{var}(Y|X = 0.5)$. *Hint: the variance of the* Uniform$[a, b]$ *distribution is given by* $(b - a)^2/12$.

(c) **[3 POINTS]** Use `numpy`'s random number generator to construct a simulated dataset of 100 samples drawn from the data-generating process above. For each sample,

- there should be one predictor feature X and one response Y
- X should be drawn uniformly from the interval $[0, 1]$
- Y should be drawn from Uniform$[X + \cos(2\pi X) - .1, X + \cos(2\pi X) + .1]$

The final dataset $\mathcal{D}$ should be stored as a data frame with two columns and 100 rows.

Use `LinearRegression` (from scikit-learn) to estimate the regression function using ordinary least squares on this dataset.

Construct a scatter plot of X and Y for your simulated dataset. On the same axes, plot the regression function $\hat{f}$ fitted by ordinary least squares. Also on the same axes, plot the true regression function f. As always, include horizontal axis label, vertical axis label, and a title. In this case, you should also include a legend (because we will have three objects plotted on the same pair of axes: the scatter plot, the estimated regression function, and the true regression function).

Report the value of $\hat{f}(0.5; \mathcal{D})$.

(d) **[3 POINTS]** Now construct 500 simulated datasets, each with 100 samples. Denote the datasets by $\{\mathcal{D}_1, \ldots \mathcal{D}_{500}\}$. For each dataset, $\mathcal{D}_i$, fit a new `LinearRegression` model and calculate $\hat{f}(0.5, \mathcal{D}_i)$. Store these estimates in a 500-dimensional vector,

$$\texttt{preds} = (\hat{f}(0.5, \mathcal{D}_1), \ldots, \hat{f}(0.5, \mathcal{D}_{500})).$$

Plot a histogram of `preds`.

(e) **[3 POINTS]** Compare the mean value of `preds` to the value you computed in part (a) of this problem. What does this suggest about the bias of the ordinary least squares estimator $\hat{f}(0.5, \mathcal{D})$?

(f) **[3 POINTS]** Compute the variance of `preds`. What does this say about the variance of the ordinary least squares estimator $\hat{f}(0.5, \mathcal{D})$?

(g) **[3 POINTS]** Suppose we gathered a new dataset with 100 samples and fit a linear regression model. Now imagine we observe a new sample with $X = 0.5$ and use our model to predict Y. Estimate the expected value of the squared error of our prediction. Note: the expectation should average over randomness in the training dataset as well as randomness in the test sample (X, Y). *(Hint: the answer can be expressed in terms of three quantities, and we have already investigated all three quantities earlier in this problem)*

Warning #1: Bias and variance depend on four things: what estimator we use, the number of samples in our dataset, the true underlying process that generates the samples, and the value of x we consider. The conclusions drawn in this problem about bias and variance are only valid for least squares estimators applied to datasets of size 100 generated by the process above to make predictions for $x = 0.5$. If you change the estimator, the dataset size, the distribution, or the value of x, the bias and variance will typically change.

Warning #2: The estimates of bias and variance in this problem were calculated using averages from 500 simulated datasets. Such simulation-based estimates are called Monte Carlo estimates. These estimates are never exactly right. However, as the number of simulated datasets increases, they typically get closer to the truth.

Total points for this assignment: 30