USC Viterbi
School of Engineering
Daniel J. Epstein Department of Industrial and Systems Engineering

ISE529 Predictive Analytics
Instructor: Dr. Tao Ma
2025 Spring

Homework 1
Due by: Feb. 7, 2025, 11:59 PM

Instructions:

1. Print your First and Last name and NetID on your answer sheets
2. Submit all your answers including Python scripts and report in a single Jupyter Lab file (.ipynb) or along with a single PDF to Brightspace by due date. No other file formats will be graded. No late submission will be accepted.
3. Total 5 questions. Total points: 100

---

1. (20 points)
For each of parts (a) through (d), indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

(a) The sample size n is extremely large, and the number of predictors p is small.

(b) The number of predictors p is extremely large, and the number of observations n is small.

(c) The relationship between the predictors and response is highly non-linear.

(d) The variance of the error terms, i.e. $\sigma^2 = Var(\varepsilon)$, is extremely high.

2. (25 points)
Answer the following questions.

(a) Provide a sketch of typical (squared) bias, model variance, training error, test error, and irreducible error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

(b) Explain why each of the five curves has the shape displayed in part (a).

(c) According to the bias-variance decomposition, show that the following equation holds.

$$E\left[f(x_0) - \hat{f}(x_0)\right]^2 = Var\left(\hat{f}(x_0)\right) + \left[Bias\left(\hat{f}(x_0)\right)\right]^2$$

3. (15 points)
Let X be the amount (in ounces) of soft drink in a randomly chosen bottle from company A, and Y be the amount of soft drink in a randomly chosen bottle from company B. A study has shown that the probability distributions of X and Y are as follows:

x	15.85	15.9	16	16.1	16.2
$P(X = x)$	0.15	0.21	0.35	0.15	0.14
$P(Y = x)$	0.14	0.05	0.64	0.08	0.09

Find $E(X)$, $E(Y)$, Var(X), and Var(Y) and interpret them.

4. (15 points)
The table below provides a training data set containing six observations, three predictors, and one qualitative response variable. Use Jupyter Lab with Python to answer the following questions.

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	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$ or 3? Why?

(c) If the Bayes decision boundary in this problem is highly nonlinear, then what would be the best choice for the value of K? Why?

5. (25 points)
Given the Auto data set (see attached **Auto.csv**), use Jupyter Lab with Python to answer the following questions. Make sure that the missing values in the data set have been removed before analysis is performed.

(a) Which of the predictors are quantitative, and which are qualitative?

(b) What is the range of each quantitative predictor?

(c) What is the mean and standard deviation of each quantitative predictor?

(d) Now remove the 10^{th} through 85^{th} observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create matrix of scatter plots highlighting the relationships among the predictors.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg?