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

ISE529 Predictive Analytics
Instructor: Dr. Tao Ma
2025 Fall

Homework 3
Due by: **March 1, 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 problems. Total points: 100

1. (20 points)
In the QDA model, it is assumed that the predictors X within each class are drawn from a normal distribution with a class-specific mean vector and a class-specific covariance matrix. We only consider the simple case where $p = 1$; i.e., there is only one predictor. Suppose that we have K classes, and that if an observation belongs to the kth class then X comes from a one-dimensional normal distribution, $X \sim N(\mu_k, \sigma_k^2)$. Show that in this case, the Bayes classifier is non-linear with quadratic form.

2. (25 points)
Use Titanic data set (**titanic_data.csv**) to complete the following tasks:

(a) Fit a **logistic regression** model to the original data with Bernoulli response format. The response variable is “Survived”; the predictors include “Pclass”, “Sex”, and “Age”.

(b) Add a new column named as “Age_Rang” to the original data. Re-label the age of each individual according to 5-year interval bin [0, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80], then use “Pclass”, “Sex”, and “Age_Range” as predictors to fit a logistic regression model where “Survived” is the response variable. Check whether or not the logistic model is the same as the one in part (a)?

(c) Transform the data set in part (b) to Binomial response format by grouping the columns “Pclass”, “Sex”, and “Age_Rang”. Fit a logistic regression to the grouped data. Check whether or not the logistic model is the same as the one in part (a) or (b)? (Hint: the response variables should include both “Survived” and “Died” columns).

(d) Are the logistic regression models in part (a), (b), and (c) adequate?

3. (20 points)
Entering high school students make program choices among general program, vocational program, and academic program. Their choice might be modeled using their writing score and their social economic status, etc. The data set contains variables on 200 students. The response variable is *prog* that is program type. The predictors include *female, race, ses* (social economic status), *schtyp* (school type), *read, write, math, science, socst* (social studies). The data can be found from ***highschooldata.csv***. Split the dataset into two subsets, i.e., the first 190 observations as the training data, and the last 10 observations as the test data.

(a) Estimate a **multinomial logistic regression** model with the training data. (prog = 1, "academic" as the baseline model).

(b) Perform prediction with the test data. Show the predicted probability as well as the program chosen.

4. (20 points)
Develop a model to predict whether a given car gets high or low gas mileage based on the **Auto.csv** data set.

(a) Add a column *mpg*01 to the original data frame. The *mpg*01 is a binary variable that contains a **1** if *mpg* contains a value above its median, and a **0** if *mpg* contains a value below its median.

(b) Explore the data graphically with matrix of scatterplots and variance-covariance matrix to investigate the association between *mpg*01 and the other features. Choose three (3) the other features that seem most likely to be useful in predicting *mpg*01?

(c) Split the data into a training set with 85% of data points and a test set with 15% of data points.

(d) Perform LDA, QDA, logistic regression, naive Bayes, and *KNN*, respectively, on the training data to predict *mpg*01 using the variables that seemed most associated with *mpg*01 in (b). What is the test error of the model obtained? Which value of *K* in *KNN* models seems to perform the best on this data set?

5. (15 points)
The number of awards earned by students at one high school follows Poisson distribution. The “num_awards” is the response variable and indicates the number of awards earned by students at a high school in a year; “math” is a continuous predictor and represents students’ scores on their math final exam, and “prog” is a categorical predictor with three levels indicating the type of program in which the students were enrolled (coded as 1 = “General”, 2 = “Academic” and 3 = “Vocational”). The data can be found in ***Awards_data.csv***. Split the dataset into two subsets, i.e., the first 190 observations as the training data, and the last 10 observations as the test data.

(a) Estimate a **Poisson** regression model with the data set.

(b) Predict the number of awards earned by the students in a year with the test data. Show the test MSE.