STSCI 3740/5740 Machine Learning and Data Mining **Fall 2024**

Dr. Nayel Bettache **Homework 3, due December 5, 11:59pm**

Problem 1 (10 points)

In the lectures, we used logistic regression to predict the probability of **default** using **income** and **balance** on the **Default** data set, which is contained in the package **ISLR**. You may use

```
library(ISLR)
data=Default
```

In this problem, you will estimate the corresponding test error based on the validation set approach and cross-validation.

1. Fit a logistic regression model that uses **income** and **balance** to predict **default**.
2. Use a validation set approach to estimate the test error in this model. To do so, consider the following steps
   (a) Use set.seed(1) to make results reproducible.
   (b) Randomly split the sample into a training set and a validation set of equal size.
   (c) Fit a multiple logistic regression of **default** with predictors **income** and **balance** using only the training observations.
   (d) Predict the default status of the persons in the validation set using a cutoff value of 0.5.
   (e) Compute the validation set error which is the fraction of individuals in the test set whose default status is misclassified.
3. Repeat the steps in part 2 using set.seed(2) and set.seed(3). Comment on the results.
4. Next, consider a logistic regression model that predicts the probability of default using **income**, **balance**, and a dummy variable for **student**. Estimate the test error for this model using the validation set approach with three seeds, **set.seed(1)**, **set.seed(2)** and **set.seed(3)**. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.
5. Similar to 4, we consider the same logistic regression model with predictors **income**, **balance**, and a dummy variable for **student**. Now, we use 5-fold cross-validation to estimate the test error rate. Comment on whether or not including a dummy variable for student leads to a reduction in the test error rate.

Problem 2 (16 points)

Solve Problem 9 on page 223 in the textbook “Introduction to Statistical Learning” (second edition).

Problem 3 (14 points)

Solve Problem 1 on page 282 in the textbook “Introduction to Statistical Learning” (second edition).

Problem 4 (4 points)

Consider the following elastic-net problem (which is another extension of lasso)

$$\min_{\beta} \sum_{i=1}^{n} (Y_i - \beta X_i)^2 + \lambda(\alpha\beta^2 + (1-\alpha)|\beta|),$$

where λ and α are two tuning parameters. For simplicity, we only consider the case that the variable X is univariate, and there is no intercept. Show how one can turn this into a lasso problem using a transformed version of X and Y.