Assignment 2

CS 750/850 Machine Learning

- **Due**: February 10th at 11:59PM
- **Submisssion**: Turn in both a **PDF** and the **source code** on MyCourses
- **Questions**: Piazza

Problem 1 [30%]

This problem examines the use and assumptions of LDA and QDA. We will be using the dataset `Default` from ISLR.

1. Split the data into a training set (70%) and a test set (30%). Then compare the classification error of LDA, QDA, and logistic regression when predicting `default` as a function of features of your choice. Which method appears to work best?
2. Report the confusion table for each classification method. Make sure to label which dimension is the predicted class and which one is the true class. What do you observe?
3. Are the LDA assumptions satisfied when predicting `default` as a function of `balance` *only* (i.e `default ~ balance`)? You can use `qqnorm` and `qqline` to examine whether the conditional class distributions are normally distributed. Also examine standard deviations of the class distributions. Are the QDA assumptions satisfied?
4. Would you ever want to use LDA in place of QDA even when you suspect that some of the assumptions are violated (e.g. different conditional standard deviations) for LDA?

Hint: Check out TidyVerse for a collection of packages that can help with data manipulation. And see the Rstudio cheatsheets for a convenient and concise reference to the methods. This is entirely optional!

Problem 2 [30%]

Using the MNIST dataset, fit classification models in order to predict the digit **1** (vs all others).

1. Compare the classification error for each one of these methods:
2. Logistic regression
3. K-NN with 2 *reasonable* choices of `k`
4. LDA
5. Explore at least one transformation of the features (predictors), such as considering their combinations, and run the methods from part 1 on the data.
6. Which one of the methods works the best?

Make sure to split the data into a training set and a test set. No need to run on the entire dataset; a subsample of say 10000 datapoints is OK.

Hint: There is a file in the gitlab repository: assignments/mnist_simple.Rmd which you can use as a starting point. If you are using Python, please checkout this package. If you have trouble getting started, please do not hesitate to ask the instructor or the TAs or Piazza for help.

Problem O2 [35%]

This problem can be substituted for Problem 2 above, for up to 5 points extra credit. The better score from problems 2 and O2 will be considered.

Solve Exercises *1.11* and *1.13* in [Bishop, C. M. (2006). Pattern Recognition and Machine Learning].

Problem 3 [20%]

Logistic regression uses the logistic function to predict class probabilities:

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

This is equivalent to assuming a linear model for the prediction of the *log-odds*:

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Using algebraic manipulation, *prove* that these two expressions are identical. See Section 4.3 in ISLR and equations (4.2) & (4.3) for more context.

Problem 4 [20%]

This problem examines the differences between LDA and QDA.

1. For an arbitrary training set, would you expect for LDA or QDA to work better on the *training set*?
2. If the Bayes decision boundary between the two classes is linear, would you expect LDA or QDA to work better on the *training set*? What about the *test set*?
3. As the sample size increases, do you expect the prediction accuracy of QDA with respect to LDA increase or decrease
4. *True or False*: Even if the Bayesian decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is more flexible and can model a linear decision boundary. Justify your answer.