CIS8695

**Homework II (Due: 2/11/2019, Midnight)**

**What to Turn In:**

Please write your answers and paste required results/reports in **a MS Word file**. **Note: Please do NOT submit any other file format, because it will cause grading inconveniency. Failing to submit in correct file format will cause the loss of homework grades! You can use screenshots if it is convenient for you.**

# Problem 1 (Ensemble Methods): Spam Emails

The file **spambase.csv** contains information on potential spam emails collected from Hewlett-Packard. Out of 4601 emails, 1813 are tagged as *spam* emails (spam=1). The predictors include 57 attributes, most of them are the average number of times a certain word (e.g., mail, George) or symble (e.g., #, !) appears in the email. A few predictors are related to the number and length of capitalized words.

1. To reduce the number of predictors to a manageable size, examine how each predictor differs between the *spam* (spam=1) and *nonspam* (spam=0) emails by comparing the spam-class average and non-spam class average. There are 11 predictors that appear to vary the most between *spam* and *nonspam* emails. Create a data frame with only these 11 predictors and the “spam” column. Partition the data with 60% training and 40% validation.
2. Fit models to the data for (1) logistic regression, (2) k-nearest neighbors with k=3, and (3) support vector classifier (linear kernel). Use “spam” as the outcome variable. Report the validation confusion matrix for each of the three models.
3. Use two ensemble methods: (1) majority voting of predicted outcomes, and (2) averaging the predicted probabilities. Report the confusion matrix for each of the two ensemble methods.
4. Compare the overall accuracy for the three individual models and the two ensemble methods, and conclude which model is the best model.

**Problem 2**. **Dimension Reduction**

Use the original dataset **spambase.csv** in above Problem 1. This time use PCA to reduce the number of predictors.

1. Using PCA, how many PCs do you think are appropriate to be used for prediction? What is the total proportion of variance do they capture?

2. Using the PCs you generate and the outcome variable “spam”, redo the step 1-4 of the above Problem 1. Report the model outputs (confusion metrics) and compare the overall accuracy of all the individual and ensemble models.