STA 4365 Final

due Tuesday May 3 by 11:59PM on Webcourses

Submission Format: Please submit your final in either Jupyter Notebook or R Markdown format. You can use at most two files for each problem (one in Jupyter Notebook and one in R Markdown). If you want, you can include multiple problems in the same file.

Problem 1: Supervised Learning (25 points) In this problem, you will investigate the stock market dataset `Weekly` included with the ISLRv2 textbook. You can load the data by using the command `library(ISLR2)` (you might need to use `install.packages(ISLR2)` if this is the first time you have used this library) and then simply calling the variable `Weekly`, which is a dataframe with the data. The response variable is `Direction`, which is a binary variable indicating whether the stock moved up or down from its previous weekly closing price at the end of the week. The predictor columns give the year, volume which indicates the number of trades, and lag variables which indicate the change in price over the lag time period. For example, `Lag2` gives the change in price over the week that occurred two weeks prior to the current week.

Prepare the data by splitting into training and validation sets. Scale the columns of the training data to have mean 0 and variance 1. Scale the validation features using the mean and standard deviation of the *training data*.

Learn the following models to predict the binary outcome:

- Naive Bayes
- Support Vector Machine with Linear Kernel
- Random Forest
- Gradient Boosted Decision Tree

For each model (except Naive Bayes), report the final hyper-parameters that you decide to use. You can either use cross-validation to tune the hyper-parameters or tune by examining the accuracy on the validation set. Train each model on the training data. Evaluate each model on testing data by reporting the accuracy and plotting an ROC curve for each model using the validation data.

Problem 2: Unsupervised Learning (25 points) This problem will examine the gene dataset from the ISLRv2 textbook in the file `gene_data.csv`. The data provides gene expression levels for 1000 different genes and 40 individuals. The first 20 observations correspond to healthy patients and the last 20 observations correspond to unhealthy patients. In this problem, you will perform unsupervised clustering to determine if unsupervised methods are able to identify the difference between unhealthy and healthy patients (the models in this part will not use information about patient health, only the gene expression levels). You do not need to split into a training and validation set for this problem.

(a) Load the data using `read.csv`. Note that this dataset is transposed: the individual observations are along the columns, and the rows give the different gene features. To correctly prepare the data, you will need to use the `header=FALSE` argument in `read.csv` since there are no column names included, then tranpose the data (switch rows and columns) using the command

```
data = t(data)
```

Prepare the data by standardizing each feature to have mean 0 and variance 1.

(b) Learn a K-means model using $k = 2$ clusters. Report the proportion of healthy and unhealthy patients in each cluster. Were the unsupervised methods able to distinguish between the two groups without knowing any labels?

(c) Visualize your K-means model by making two 2D plots of the data: one using the first two principal components from PCA, and another using t-SNE. You can use the same t-SNE parameters that were used in HW3 (perplexity of 30 and PCA pre-processing using 50 dimensions). Make a plot for each visualization method and color-code each observation according to the k-means labels.