STA 4365 Midterm

due Monday March 28 by 11:59PM on Webcourses

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

Problem 1: (25 points) In this problem, you will investigate the income dataset from the file `adult.data`. The response variable is the last column, which is a binary variable that indicates whether an individual earned greater than or less than $50K in a year. Information about the attributes can be found here. Load the data using `read_table` or `read.table`. Format all categorical data correctly. Split your data into a training and validation set. Scale the columns of the training and testing features using the mean and standard deviation of the *training data*.

Learn the following models to predict the binary outcome:

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

For each model, report the final hyper-parameters that you decide to use. You can either use cross-validation to tune the hyper-parameters or tune by hand 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: (25 points) This problem will examine the Parkinson's Dataset which is described at this site. The intended goal of this dataset is to predict the `status` variable, which is a 0/1 indicator variable that describes if a patient has Parkinson's based on attributes of speech patterns. You can remove the `name` features. In this problem, you will perform unsupervised clustering on the predictor features (all features excluding `name` and `status`). You do not need to split into a training and validation set for this problem.

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

(b) Next, you will learn a K-means model on your unsupervised features. Plot the silhouette score across the different number of clusters K from 2 to 15. What is the optimal number of clusters for this dataset for each model? Learn a final K-means model using the optimal number of clusters.

(c) Suppose you have observed a new observation X_{new} where $X_{\text{new}}^{(j)} = \frac{1}{n}\sum_{i=1}^{n} X_i^{(j)}$ (in other words, X_{new} is just the mean vector across the observed features). Find the cluster that X_{new} belongs to by finding

$$k^*_{\text{k-means}} = \text{argmin}_k \|X_{\text{new}} - \bar{X}_k\|_2^2.$$

(d) Now suppose that you obtain the data labels in `status` after you have learned your clusters. Find the number of observations with and without Parkinson's in each cluster. Use a χ^2 test to determine if cluster membership is independent of Parkinson's status. If they are not independent, describe the relationship between cluster membership and Parkinson's status.