SEIS 763 Machine Learning
Assignment 3
Due: midnight 10/27/20 on Canvas

Individual effort

You will be implementing classification models on the dataset provided.

Dataset:

- A telecommunications provider has categorized its customer base by service usage patterns into four groups. If demographic data can be used to predict group membership, the company can customize offers for individual prospective customers. It is a classification problem. That is, given the dataset, with predefined labels, we need to build a model to be used to predict class of a new or unknown case.
- We will use demographic data, such as region, age, marital, etc. to predict usage patterns.
- The target field, called **custcat**, has four possible values that correspond to the four customer groups, as follows: 1- Basic Service, 2- E-Service, 3- Plus Service, 4- Total Service.
- Our objective is to build classifiers, to predict the class of unknown cases.

What you need to do: Create a jupyter notebook called **Assign3.ipynb**. Write code for each of the following questions by having a separate cell for every question. Copy the actual question in a markdown cell and right below that you should have a code cell.

1) Display the number of instances for each class. That is, you should write code to find out how many instances you have for class 1, class 2, class 3, and class 4. (No loops)

2) Perform one hot encoding on Column 1 (i.e. region) and drop the extra dummy variable.

3) Create histograms of columns age and income to visually explore their distributions.

4) Split the dataset into training (70%) and testing set (30%). Perform normalization of the data using standardization.

5) **Model 1**: Fit a logistic regression model. What is the testing misclassification rate you get?

6) **Model 2**: We will now fit k-NN. However for k-NN you need to specify the value for k. In order to figure that out, run k-NN in a loop with different values of k (starting from k=5) and compute the testing misclassification rate. Plot a chart with k on X-axis and testing error on the Y-axis. What is the lowest value of testing error and corresponding value of k?

7) **Model 3**: Fit SVM model with different kernels. Which kernel gives the least testing error?

8) **Model 4**: Fit Naïve Bayes model. What is the testing error you get?

9) **Model 5**: Fit Random Forest model. For Random Forest, you need to specify the number of trees (n_estimators). In order to figure that out, run Random Forest in a loop with different values of n_estimators (starting from 10) and compute the testing misclassification rate. Plot a chart with n_estimators on X-axis and testing error on the Y-axis. What is the lowest value of testing error and corresponding value of n_estimators?

Q10) **Predicting with Ensemble:** Now that you have built 5 models. Loop over the testing set. For every test instance, have each of the models predict the class label. Eventual class predicted will be based on a majority vote of the 5 models. What is the testing misclassification rate you get with the ensemble model?

Submission:

- Make sure each of the cells have been run along with the output shown right below. Now, export the notebook as .html file.
- Submit the **.html** file and **.ipynb** notebook on Canvas.

Note: Do not submit the data. Your code should be referencing the data file when you load the data in your code.