Assignment 3

Due: March 18

Submission Instructions

- **Create a README file, with simple, clear instructions on how to compile and run your code. *If the TA cannot run your program by following the instructions, you will receive 50% of programing score.***
- **Zip all your files (code, README, written answers, etc.) in a zip file named** $\{firstname\}\_\{lastname\}\_CS5790\_HW3.zip$ **and upload it to Blackboard**

1. (10 points) Sign up for project teams in this shared document:

   `https://docs.google.com/a/fordham.edu/spreadsheets/d/1M3klIMqPSoIQKw9-chzd9hgsbupxkTw1gT9aq7bbvnE/edit?usp=sharing`

   Each team should have 1- 4 members. Please fill in the last two columns indicating if your team is open for more members and what programming language(s) your team is planning to use.

For Question 2 and 3, you will be extending your KNN classifier to include automated feature selection. Feature selection is used to remove irrelevant or correlated features in order to improve classification performance. You will be performing feature selection on a variant of the UCI vehicle dataset in the file `veh-prime.arff`. You will be comparing 2 different feature selection methods: the Filter method which doesn't make use of cross-validation performance and the Wrapper method which does.

Fix k = 7 for all runs of LOOCV in Question 2 and 3.

2. (30 points) Filter Method
   Make the class labels numeric (set "noncar"=0 and "car"=1) and calculate the Pearson Correlation Coefficient (PCC) of each feature with the numeric class label. The PCC value is commonly referred to as r. For a simple method to calculate the PCC that is both computationally efficient and numerically stable, see the pseudo code in the `pearson.html` file.

   (a) List the features from highest $|r|$ (the absolute value of r) to lowest, along with their $|r|$ values. Why would one be interested in the absolute value of r rather than the raw value?

(b) Select the features that have the highest m values of $|r|$, and run LOOCV on the dataset restricted to only those m features. Which value of m gives the highest LOOCV classification accuracy, and what is the value of this optimal accuracy?

3. (40 points) Wrapper Method
Starting with the empty set of features, use a greedy approach to add the single feature that improves performance by the largest amount when added to the feature set. This is Sequential Forward Selection. Define performance as the LOOCV classification accuracy of the KNN classifier using only the features in the selection set (including the "candidate" feature). Stop adding features only when there is no candidate that when added to the selection set increases the LOOCV accuracy.

   (a) Show the set of selected features at each step, as it grows from size zero to its final size (increasing in size by exactly one feature at each step).

   (b) What is the LOOCV accuracy over the final set of selected features?

4. (20 points) Ensemble Learning

   Suppose we need to build a predictive model for a binary classification task. We have 25 students in our class. Each of us worked independently and everyone is able to build a model with 60% accuracy.

   (a) If we take 3 models and build a majority vote classifier C_3, what would be the accuracy of our new classifier C_3? Show your work.

   (b) If we take 5 models and build a majority vote classifier C_5, what would be the accuracy of our new classifier C_5? Show your work.

   (c) If we take all 25 models and build a majority vote classifier C_{25}, what would be the accuracy of our new classifier C_{25}? Show your work. (You may need to write a small program to compute this).

   (d) The performance you obtained for C_{25} is too good to be true. What's the assumption in your calculations that often does not hold in reality?

   (e) What would be the answer to (c) if everyone's model only has 45% accuracy? Show your work.