Assignment 3

Due: March 21

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. (25 points) Decision Tree

   Table 1 below contains a small training set. Each line includes an individual's education, occupation choice, years of experience, and an indication of salary. Your task is to create a complete decision tree including the number of low's & high's , entropy at each step and the information gain for each feature examined at each node in the tree.

Instance	Education Level	Career	Years of Experience	Salary
1	High School	Management	Less than 3	Low
2	High School	Management	3 to 10	Low
3	College	Management	Less than 3	High
4	College	Service	More than 10	Low
5	High School	Service	3 to 10	Low
6	College	Service	3 to 10	High
7	College	Management	More than 10	High
8	College	Service	Less than 3	Low
9	High School	Management	More than 10	High
10	High School	Service	More than 10	Low

Table 1: Decision Tree Training Data

Please turn in a diagram similar to:

Top 6,4, .97
Education gain = <to be calculated>
1. High School 4,1, <to be calculated>
   Experience gain = <to be calculated>
   Etc.
Etc.

Prune the tree you obtained using the validation data given in Table 2. Show your work.

Instance	Education Level	Career	Years of Experience	Salary
1	High School	Management	More than 10	High
2	College	Management	Less than 3	Low
3	College	Service	3 to 10	Low

Table 2: Validation Data

2. (25 points) Build a Naive Bayes classifier using the training data in Table 1 with **add 1 smoothing** technique covered in the lecture slides. Use your model to classify the following new instances:

Instance	Education Level	Career	Years of Experience
1	High School	Service	Less than 3
2	College	Retail	Less than 3
3	Graduate	Service	3 to 10

For Question 3 and 4, 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 the KNN parameter to be k = 7 for all runs of LOOCV in Question 3 and 4.

3. (20 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.

   (1) 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?

(2) 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?

4. (30 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.

   (1) 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).

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