CSCI 7350/6600 Data Mining
Project 2 (due midnight, Tuesday, Oct 13th)

In this project, we experimentally study the following classification schemes: (1) logistic regression, (2) K nearest neighbor, (3) decision tree classification, (4) Naive Bayes classification, and (5) Artificial Neural Network using scikit learn (https://scikit-learn.org/stable/) .

Data

We will use a data set from the UCI machine learning repository at https://archive.ics.uci.edu/ml/datasets.php. Follow the link to the repository website and select the data sets under the tab "Classification tasks". The description page of each data discusses the past usage of the data, the attributes and the valid values of the attributes used in data, and the characteristics of the data, e.g., number of examples, whether there are missing attribute values, etc. The data set to be used is the "Heart Failure clinical records Data Set".

Experiment and Analysis

A. Logistic Regression

Apply 10-fold cross validation to compute the average prediction accuracy on the data.

B. K Nearest Neighbor

Apply 10-fold cross validation to find the best K value for the data set. Experiment with potential best K value in the range [1, 15] (odd numbers only). Plot the prediction accuracy vs. K value plot to assist in deciding on a good K value to use for this data. Report the average prediction accuracy.

C. Decision Tree

1. **Display the decision tree derived from the data. If the decision tree is too big, you can set the max_depth to a reasonable depth**
    - Show two complete decision rules that can be used to predict each of the classes

2. **Compare the cross-validation results of the decision tree classification using the entropy vs. the gini index criterion function**

Decision Tree	Average Prediction Accuracy
criterion = entropy	
criterion = gini index	

D. Naïve Bayes

Most data sets have a combination of both numeric and nominal valued attributes. Depending on whether you choose to use Gaussian NB, or the Multinomial NB, you may need to convert all the attribute values into numeric, or to discretize the numeric data into nominal attribute as a pre-processing step.

Another option for working with data with mixed numeric and categorical data is to separate the data into two parts, one with numeric valued attributes and one with categorical data. Run Gaussian NB and Multinomial NB on each separably and compute the final posterior probability by multiplying the results from the two classifiers. classifier.predict_proba can be used to display the posterior probability value from a classifier.

E. Artificial Neural Network

Experiment with hyper-parameters used in NN classification. (use training-test data 75%, 25% split for this experiment

1. Experiment with number of hidden layers used in the NN structure, vary the number of layers from 1 to 5. Compare the test accuracy results. (Set the batch size to 20),
2. Select the batch size. Experiment with setting the batch size to 10, 20, 30 and 40 and Compare the test data accuracy results. (Use the number of hidden layers identified to give the best average accuracy for this experiment),
3. Experiment with the learning rate parameter to find a good learning rate to use for this data.

Data Preprocessing

To prepare the data for classification, lookup the sklearn preprocessing API docs, for example:

- sklearn.preprocessing.LabelEncoder
- sklearn.preprocessing.StandardScaler
- sklearn.preprocessing.OneHotEncoder

Things to turn in:

- Turn in the report, programs, and data in D2L Dropbox under "Project 2":
    - Write a mini-report that contain:
        - A brief description of each of the classification method;
        - A description of the data set used in the experiments;
        - Experimental steps taken and the results (plots) obtained, and
        - Analysis/discussion of these results;
        - A conclusion from these experiments.
    - The python programs that run the experiments, and
    - The data files used