CSE 487/587 Assignment 1:Predictive Analytics

The objective of this assignment is to get started with Predictive Analytics. The goals of the assessment is to implement a predictive analytics algorithm from scratch and to create a Scikit-learn predictive analytics pipeline and perform visualization using Matplotlib.

PART - 1

Implementation of Machine Learning Algorithms

- Implement Supervised Learning algorithms K-NearestNeighbour and Random Forest
- Implement Dimensionality Reduction using PCA
- Implement K-Means Clustering
- Implement Evaluation measure Accuracy, Recall, Precision, Within cluster Sum of squares

The implementation should only use NumPy and Pandas as external libraries. The implementation should be within the definition of the functions in the predictive_analytics.py file. Any utility function should be defined within its scope.

Evaluation - Total of 55 Points

- Function implementing K-NearestNeighbour algorithm (5 Points)
- Function implementing Random Forest (25 Points)
- Function implementing PCA (5 Points)
- Function implementing K-Means Clustering (12 Points)
- Function for Accuracy (2 points)
- Function for Recall (2 points)
- Function for Precision (2 points)
- Function for Within Cluster sum of squares (2 points)

The points are for the correct implementation. This is validated by getting comparable accuracy on the test set with the standard library like Scikit-learn. Make sure that the code is optimized and does not take a long time to run.

PART - 2

Scikit-Learn Pipeline for Machine Learning

- Using the Scikit-learn library, implement supervised learning algorithms SVM, Logistic Regression, Decision tree, KNN
- Using the Scikit-learn library create an ensemble model using the voting classifier of the above-mentioned algorithm.
- Create visualization using Matplotlib library, of the confusion matrix for each of the 5 classifiers (including the ensemble).

Evaluation:- Total of 45 Points

- Implementation of supervised algorithms SVM, Logistic Regression, Decision Tree, KNN, using Scikit-learn and report accuracy (10 Points)
- Creating an ensemble model using the voting classifier of the above-mentioned algorithms and report accuracy (10 Points)
- Creating visualizations using Matplotlib library for the confusion matrix of each model (use subplot) (10 Points)
- Perform Grid search on hyperparameters of SVM, Decision Tree and KNN and create plots using Matplotlib (15 Points)

Dataset:-

You are provided with data.csv along with this document. The data has 48 features and the last column corresponds to the label. Use this to implement your algorithms. Your algorithms will be evaluated on a private test set, and points will be allocated on the basis of the performance on this test set.

Submission Instructions

You will submit Assignment1.zip or Assigment1.tar.gz, a compressed archive file containing the following files:

- Predictive_analytics.py file
- The figure for confusion matrix visualization
- One plot each for SVM and Decision Tree reporting hyperparameter search
- Text file with your and your teammates' names and UB ID (only one submission per team)

Submission is due 03/09/2020,Monday, 11:59 PM EST. Please use the submit_cse487 or submit_cse587 script in Timberlake to submit your assignment.