Assignment 2

NetID: zt414 Name: Zhaoqing Teng

In this assignment, I implemented three supervised learning algorithms (k-nearest neighbor, perceptron and ID3 decision tree learning) and two unsupervised learning algorithms (k-means and Agglomerative Nesting).
The first file is OtherTool.java. In this file I implement a new data structure stateoverview to store the information I get from the file, together with the basic function to initiate or adjust the data structure . I also implement the readFile function to read the data from the parameter “path”.
The second file is K-Nearest-Neighbor algorithm. The idea is like the slices. I stored the information of the train data and tested the test data to see which K data pieces in the train data have the shortest distance to the test data. The K is the parameter and I found that 20 is best according to the performance. I also add a vector of weight in my program because I believe different element should have different influence on the result. And the result I get from this algorithm is above 92%
The third file is perceptron algorithm. This algorithm is executed according to the weight parameters I calculated from the train-data. And then use these weight parameters to calculate the result of the test data. And the result of this algorithm is above 83%.
The fourth file is ID3 decision tree algorithm. Firstly I transferred the data into 1,2,3,4 these four kinds of possible value of every parameter. And then I built the decision tree according to the value of each parameter and the information gain. Then I use this tree to test the test data and to decide the result. The result of this algorithm is above 95%
The fifth file is K-mean algorithm. This algorithm is to divide the train data into two dataset. One votes for Democratic and the other one votes for Republican. The idea is to generate K clusters randomly and then adjust the center of the cluster to involve all of the data in the train data. The result is right and decided the data into two datasets. And one of them has all of the data which voted Democratic and the other one has all of the data which voted Republican. The silhouette coefficient of this algorithm is 0.85.
The sixth file is the Agglomerative Nesting algorithm. The idea is to generate clusters every two data according to the distance. And then merge the clusters into the K clusters we want. In this assignment I also divided the data into two dataset, one voted Democratic and the other one voted Republican. The silhouette coefficient of this algorithm is 0.85.