Machine Learning Lab 8

K-Means Clustering and KNN

K-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.The algorithm starts by randomly choosing a centroid value for each cluster. After that the algorithm iteratively performs three steps:
(i) Find the Euclidean distance between each data instance and centroids of all the clusters
(ii) Assign the data instances to the cluster of the centroid with nearest distance
(iii) Calculate new centroid values based on the mean values of the coordinates of all the data instances from the corresponding cluster.

KNN is the abbreviation for K-Nearest Neighbours. KNN algorithm is one of the simplest classification algorithm and it is one of the most used learning algorithms. it is a non-parametric, lazy learning algorithm. With non-parametric we mean that it does not make any assumptions on the underlying data distribution. The output of a KNN algorithms is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its K nearest neighbors

The dataset

The dataset used to perform this experiment is the wine quality dataset, it is a combination of data on two types of wine variants, namely red wine and white wine, of the portuguese “Vinho Verde” wine. The dataset contains information on the parameters for fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol.

Experiment

In this experiment I used the sklearn’s K-Means and KNN algorithms to predict the quality of a wine.

Using the pandas library in I loaded the red wine and white wine datasets into the memory from their respective csv files and then merged the two datasets into one single pandas dataframe.

Using the pandas.Dataframe.describe() function in pandas I calculated the various statistical measures of each of the columns of the dataset.

For performing the experiment I started with scaling the features in the dataset. Since the features of the dataset each have a different scale we need to bring them to the same scale so that none of the features dominates the others features while training the models.

Next ran the K-Means algorithm on the data for the number of centroids in the range of 1 to 20 to find the the number where the elbow point occurred. As evident in the notebook it occurs at K=5 or 6, choosing K=5 I plotted the ingridients of the wine pair wise. From which we conclude that:

Cluster 1: Low pH, high sulphates, low alcohol
Cluster 2: High pH, low sulphates, high alcohol, low total sulpur dioxide
Cluster 3: Low alcohol, low sulphates, high total sulpur dioxide
Cluster 4: High alcohol, low pH, low total sulpur dioxide
Cluster 5: Low alcohol, low sulphates, low total sulphur dioxide

Thus we see that:
pH is high in cluster 2 and low in cluster 1.
Sulphates is high in cluster 1 and low in cluster 3.
Alcohol is high in cluster 2 & 4 and low in Rest of the clusters(1,3,5).
Total surfur dioxide is high in cluster 3 and low in cluster 4.

For the KNN I used a similar approach as K-Means by running the KNN algorithm over data with changing only the K value from 1 to 160 and as evident in the graph the value decrease as the K value increases. However it seems to plateau around 0.8.

The code and plots can be found in the accompanying jupyter notebook.