CO324 Pattern Recognition
Assignment-2

Siddhant Verma
2K18/EC/167

April 5, 2021

Exercise 1

Use the satellite (UCI Machine Learning Repository: Statlog (Landsat Satellite) Data Set) dataset for this question.

1. Load the dataset and perform splitting into training and validation sets with 70:30 ratio. Use tsne plot to visualise the dataset.
2. Implement the kNN algorithm from scratch. You need to find the optimal number of k using the grid search. You may use sklearn for grid search. Plot the error vs number of neighbors graph (k). Report the optimal number of neighbours.
3. Report the training and the validation accuracy only with optimal value of k using sklearn kNN function. Comment on the accuracy obtained for optimal value of k for both the methods i.e, your implementation and the inbuilt sklearn function.

Solution:

KNN Algorithm from scratch:

- function knn-distances used to calculate euclidiean distances between 2 matrices.
- knn-predictions function appending to a list of labels of all the nearest neighbours and returning the class which is present the most in the given k number of neighbours
- knn-accuracy to calculate the accuracy of classifications

Performance of KNN from scratch:

- Validation Set accuracy: 90.61% with k=1 nearest neighbours.
- Testing Set accuracy: 89.95% with k=4 nearest neighbours.

Performance of KNN from Sklearn:

- Validation Set accuracy: 90.61% with k=1 nearest neighbours.
- Testing Set accuracy: 89.7% with k=3 nearest neighbours.

The performance achieved by my implementation is almost as accurate as that of the inbuilt kNN sklearn function.

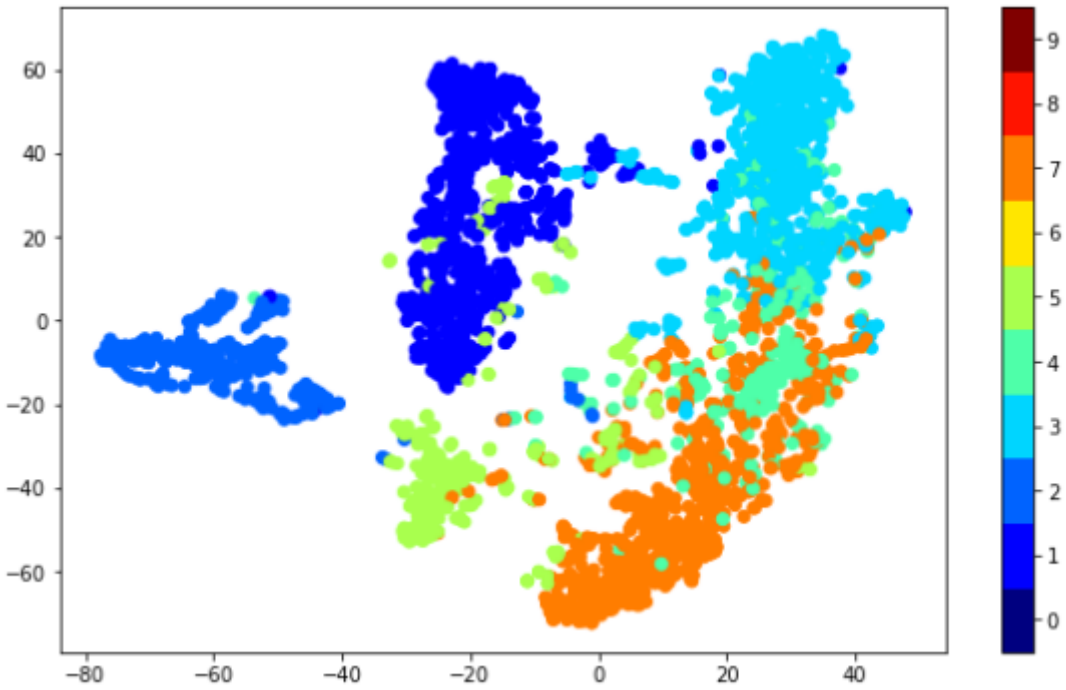

Figure 1: TSNE

Exercise 2

Use the IRIS (UCI Machine Learning Repository: Iris Data Set) dataset for this question.

1. Load the dataset and perform splitting into training and validation sets with 70:30 ratio.
2. Implement the Kmeans algorithm using sklearn. You need to find the optimal number of clusters using the elbow method. Plot the error vs number of clusters graph while using the elbow method. Report the optimal number of cluster found.
3. Use Scatter plot to visualize the dataset to depict the clusters formed (optimal).
4. Report the training and the validation accuracy. Comment on the accuracy obtained for both the sets.

Solution:
The techniques of the elbow method is to run k-means clustering on the dataset for a range of values of k (say, k from 1 to 10), and for each value of k calculate the sum of squared errors (SSE). Then, plot a line chart of the SSE for each value of k. If the line chart looks like an arm, then the "elbow" on the arm is the value of k that is the best.

Our goal is to minimize SSE, but the SSE tends to decrease towards 0 as we increase k (the SSE is 0 when k is equal to the number of data points in the dataset, because then each data point is its own cluster, and there is no error between it and the center of its cluster). So our goal is to choose a small value of k that still has a low SSE, and the elbow usually represents where we start to have diminishing returns by increasing k.

The optimal number of clusters found = 3.

Performance:

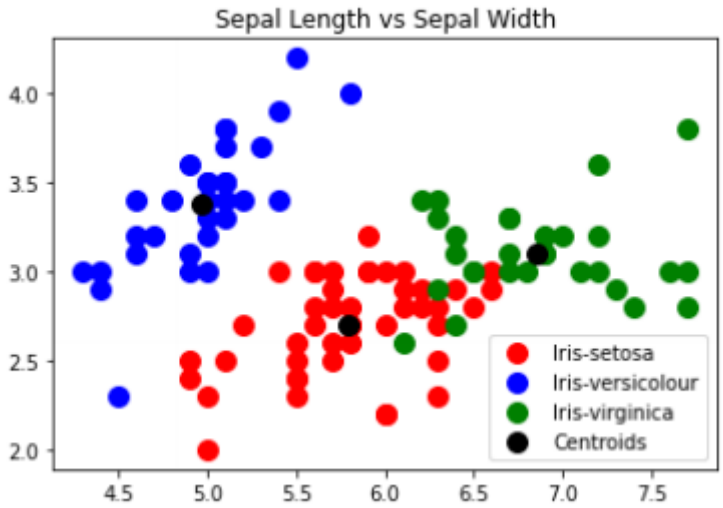

Figure 2: Scatter Plot

- Training Set Accuracy: 24.76
- Validation Test Accuracy: 93.33

This is a case of underfitting. This indicates the presence of high bias in the dataset.

Smaller datasets have smaller intrinsic variance so this means that your model properly captures patterns inside of your data and train error is greater simply because the inner variance of training set is greater then validation set. (find graphs below)

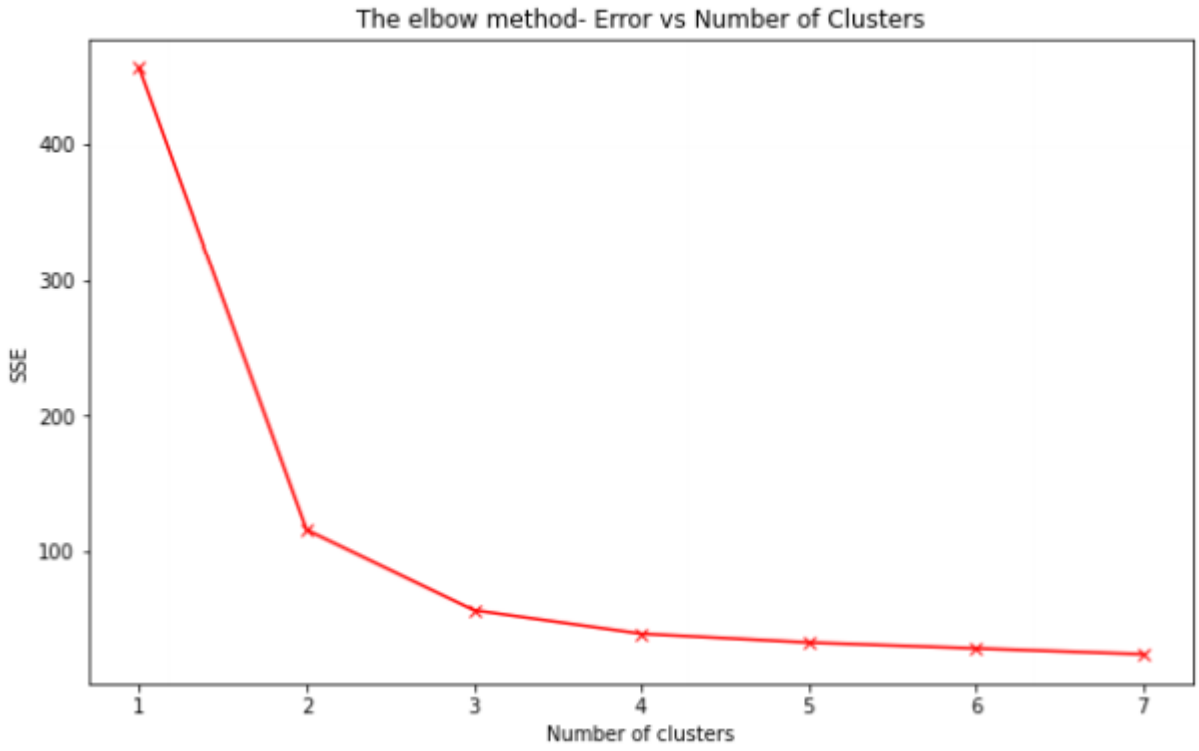

Figure 3: Kmeans using elbow method

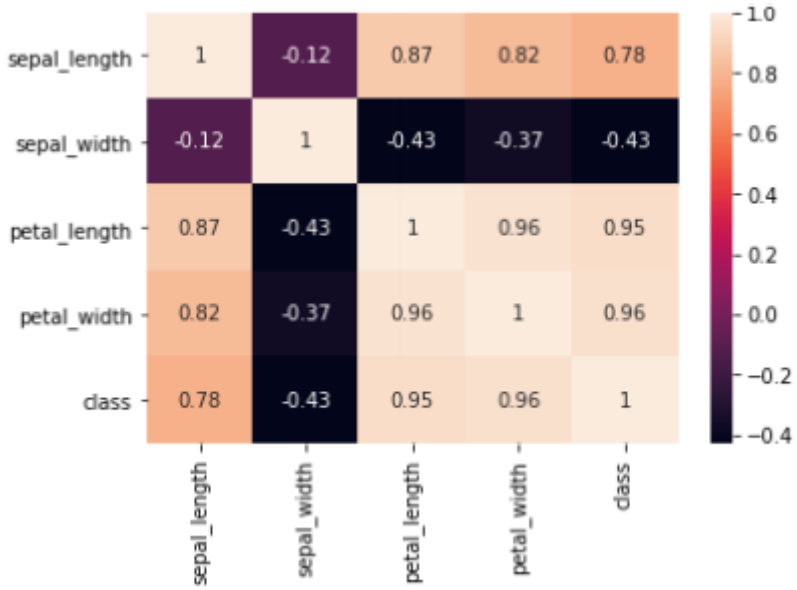

Figure 4: Autocorrelation matrix