CO544 : Machine Learning and Data Mining
Lab 05: Classification, Predictions, Clustering and Association Learning
Nawarathna K.G.I.S.
E/17/219

Part 3 : Clustering

3. Briefly describe what is meant by the term seed in the "Generic Object Editor". Describe the use of seed with the KMeans algorithm.

> Seed is used to seed the random number generator. Initially since the program has no idea where to pick points from, the seed is used to set some points as the centroids. For this main reason, seed is used.

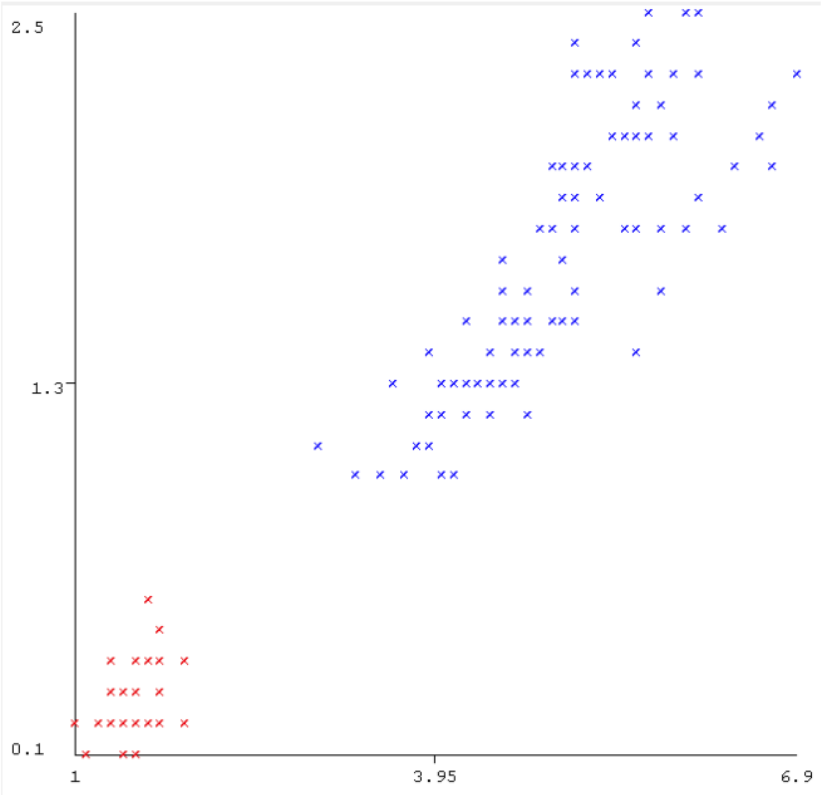
4. Observe the cluster assignments of SimpleKMeans algorithm and describe the values in each cluster. Record the sum of squared error and the proportions of instances assigned to each cluster.

Sum of squared error : 12.144
Proportions of instances assigned to each cluster:
Cluster0 : 100 (67%)
Cluster1 : 50 (33%)

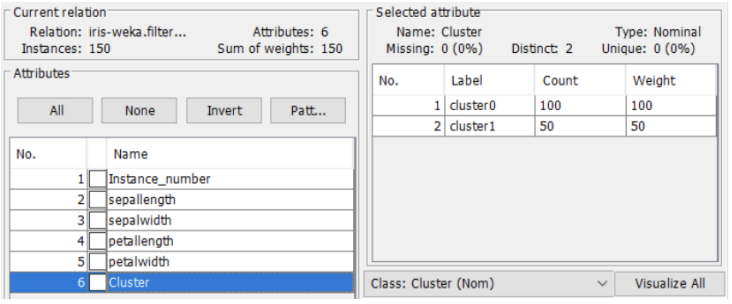
```
Final cluster centroids:
                            Cluster#
Attribute      Full Data          0          1
                 (150.0)    (100.0)     (50.0)
=============================================
sepallength       5.8433      6.262      5.006
sepalwidth         3.054      2.872      3.418
petallength       3.7587      4.906      1.464
petalwidth        1.1987      1.676      0.244
```

5. Briefly describe your observations found in Description in each cluster. Choose suitable labels for “X”, “Y”, “Colour” fields.

- The feature selection is done considering the separation of each cluster. It has made sure that clusters have separated with a clear margin.
- Labels :
    - x : petallength (Num)
    - Y : petawidth (Num)
    - Colour : cluster (Nom)

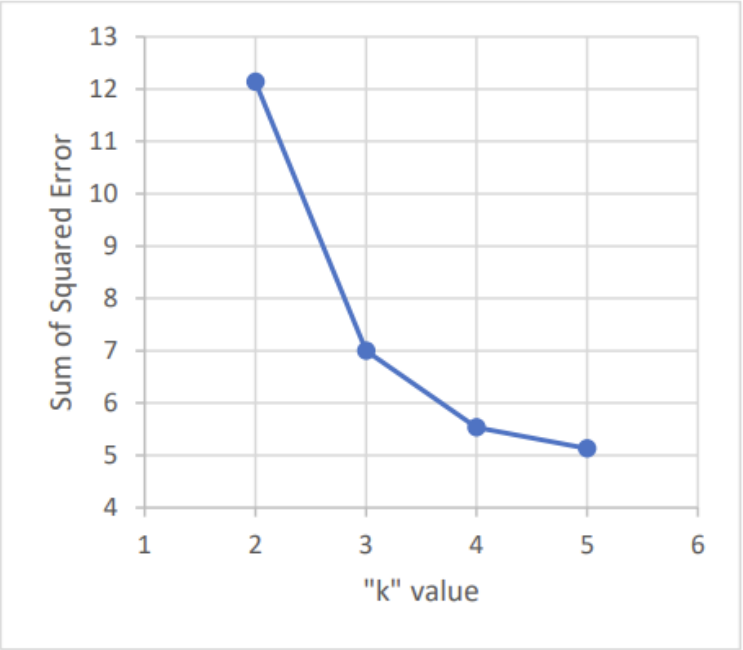


6. Briefly describe the content of the ARFF file“iris-kemeans-noofclusters-2.arff”.



Clusters themselves are also in the dataset, other than features.

7. Repeat the above process for different values of “k” ($2 \leq k \leq 5$) and suggest a suitable value for “k”. Justify your answer.



Squared error for k values :

k= 2; error = 12.14

k =3; error = 7

k=4; error = 5.53

k=5; error = 5.13

8. Evaluate the model for the “k” value you decided above and observe the results.

```
=== Model and evaluation on training set ===

Clustered Instances

0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)


Class attribute: class
Classes to Clusters:

  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances :       17.0     11.3333 %
```