Danny North

4/1/16

CS 478

Dr. Martinez

Cluster Lab

I. **Sponge Data**

The results from the Sponge Data are shown here. The algorithm converts any unknown values to 1, which seems to work and converge after 4 iterations. This is the same for all future datasets.

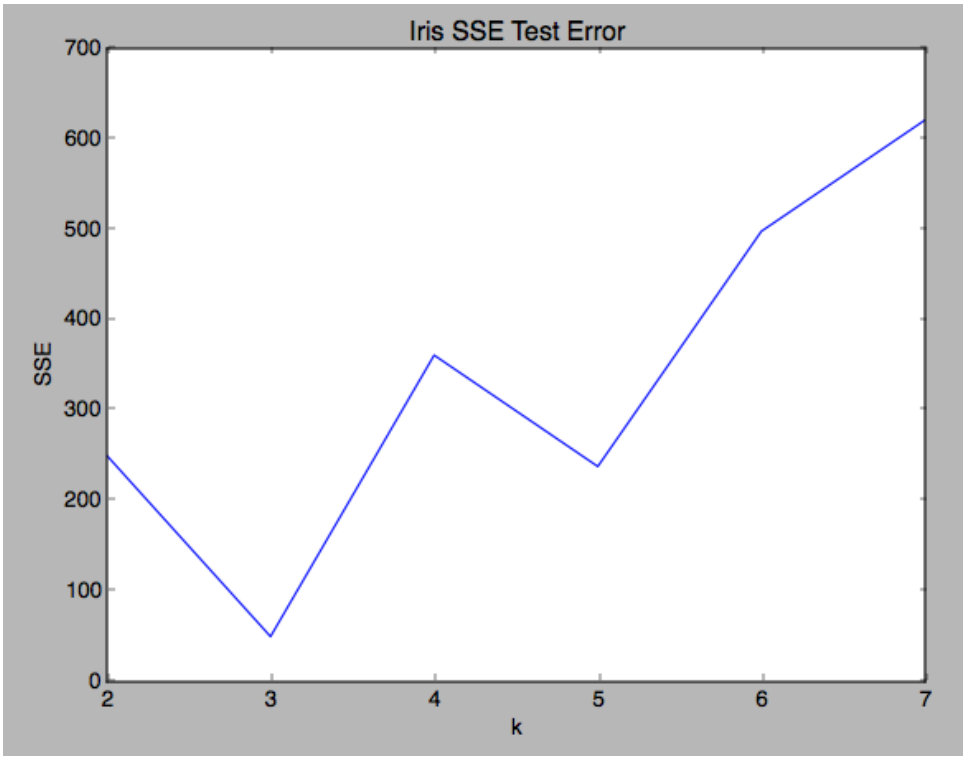
```
Iter: 0 SSE: 115.000
Iter: 1 SSE: 125.000
Iter: 2 SSE: 131.000
Iter: 3 SSE: 125.000
Iter: 4 SSE: 125.000
k: 4, SSE: 114.000
Final Centroids:
Centroid  0  =  [ 0.5 3.  1.  2.75 2.821  2.928 2.464  1.321]
Centroid  1  =  [ 2.96  3.033  0.033  0.066  4. 2.66 1.96  0. ]
Centroid  2  =  [ 3.  3.  0.  0.   4.   3.   0.5   1.25]
Centroid  3  =  [ 1.214  3.357  1. 2.928  0.928  0.571 2.857  2.357]
```

Here are the results for the normalized version of the data.

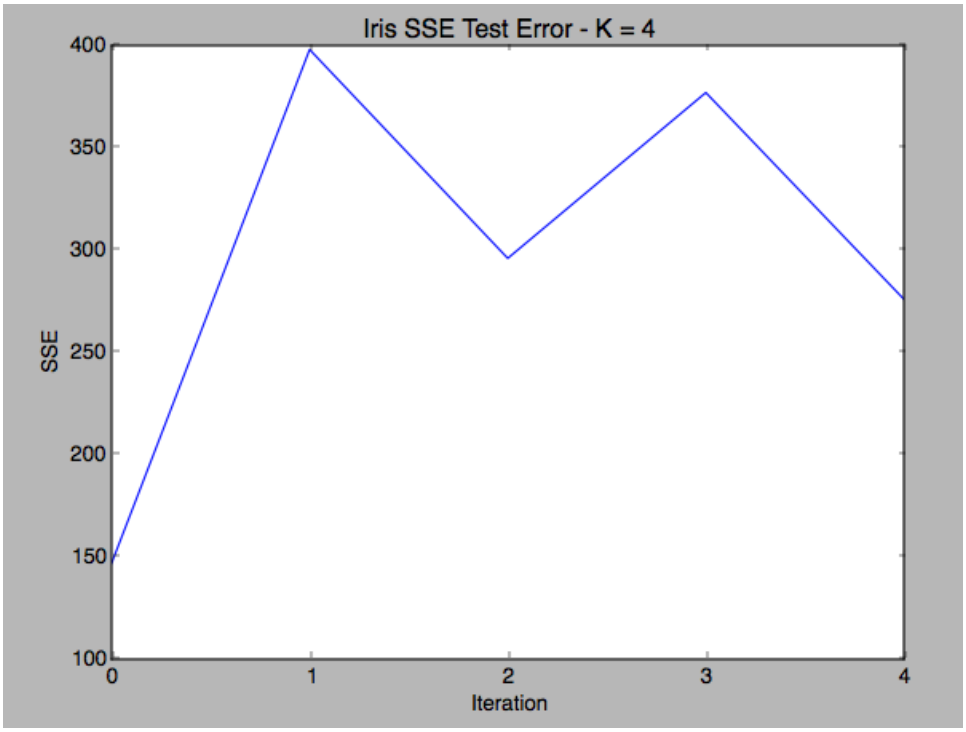
```
Iter: 0 SSE: 116.000
Iter: 1 SSE: 118.000
Iter: 2 SSE: 116.000
Iter: 3 SSE: 116.000
k: 4, SSE: 115.000
Final Centroids:
Centroid  0  =  [ 0.255  0.627  1.  0.697  0.558  0.720 0.852 0.406 ]
Centroid  1  =  [ 1.  0.6  0.  0.  1.  0.888   0.870  0.  ]
Centroid  2  =  [ 1.   0.6   0.   0.   1.  0.904  0.285  0.036 ]
Centroid  3  =  [ 1.   0.6   0.   0.   1.   1.  0.333  0.75 ]
```

II. **Iris Data**

The iris dataset is not normalized in the following tests. Below are the results for k= 2- 7 and the SSEs. There seems to be an overfit when using k = 7. The results look better for k = 3, generally. This makes sense because in reality there are 3 iris types in this dataset.



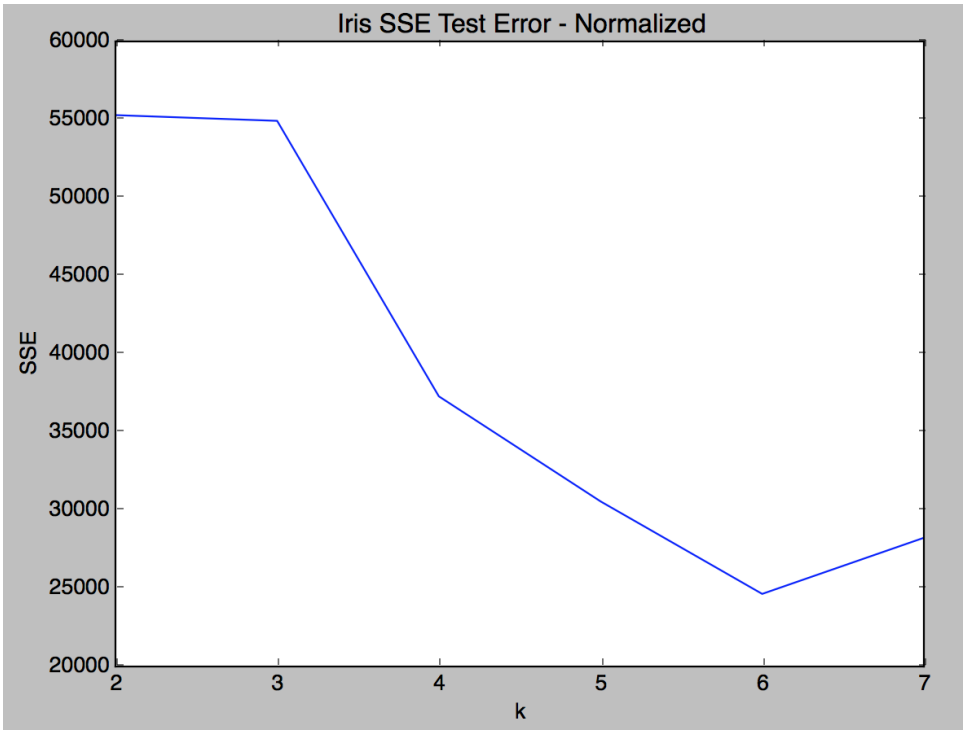
Now trying this dataset with the output as a feature. When running this five times with k=4 we get the following results. There is a little variation between the iterations because of the randomness factor. If the randomness factor is good, then there should be very little variation in the SSEs between the numbers. Although for this dataset we seem to be getting an average of 298, where the average for the dataset while not using the output as a feature was about 258. Therefore, putting the output as a feature did not actually improve our clustering.



III. Abalone Dataset

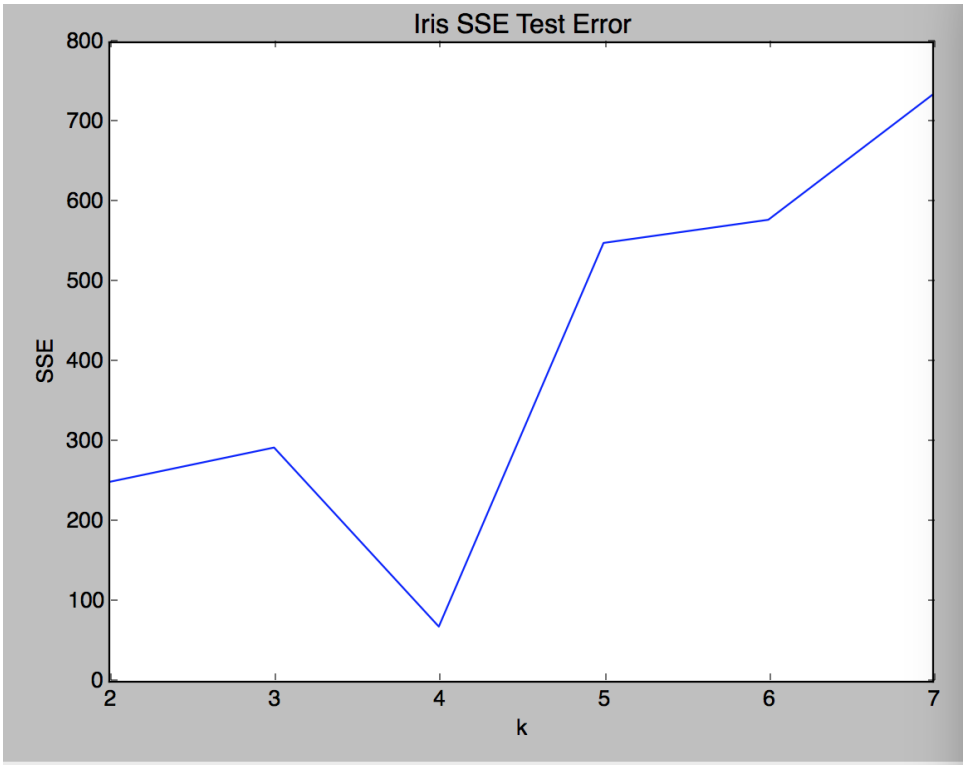
For the Abalone Dataset the results show that the SSE gets lower the more clusters we add. This makes sense because there are actually 29 different clusters, although the data seems to be centered more across 18 main outputs. It was suggested to use "Rings" as a continuous variable, probably because it is an age and makes sense to be used as a continuous value. Below are the graphs shown for both non-normalized and normalized data.


Iris SSE Test Error
60000
55000
50000
45000
40000
35000
30000
SSE
2
3
4
5
6
7
k


Iris SSE Test Error - Normalized
60000
55000
50000
45000
40000
35000
30000
25000
20000
SSE
2
3
4
5
6
7
k

IV. Experiment

For my experiment I decided to test different values for the unknowns to see what kind of results I could get for the iris dataset. Changing the value from 1 to 2 produced very little results. Changing it to 14 gave it a shape that seemed to favor 4 as its k-value. This is interesting because the previous results showed 3 or 4 were preferable, but this one makes it clear that it prefers a k-value of 4.



I also decided to test replacing unknown values with the mean along that column. This gave us the same shape that we were seeing before, but interestingly enough, didn't improve the SSE value like I thought it would. It was generally worse than just choosing 1 as a distance, although not by much.