UUN: S1525701

Task1

Task 1.2

Predicted	1	2	3	4	5	6	7	8
Actual								
1	160	78	501	623	367	85	90	96
2	419	46	223	89	345	515	278	85
3	55	729	200	50	251	52	409	254
4	312	101	375	93	572	203	184	160
5	425	271	108	61	173	63	82	817

In the confusion matrix for the train data above the rows represent the actual class labels – numbered from 1 to 5 - and columns represent the 8 classes, as labelled by the K-means clustering – labelled 1 through 8.

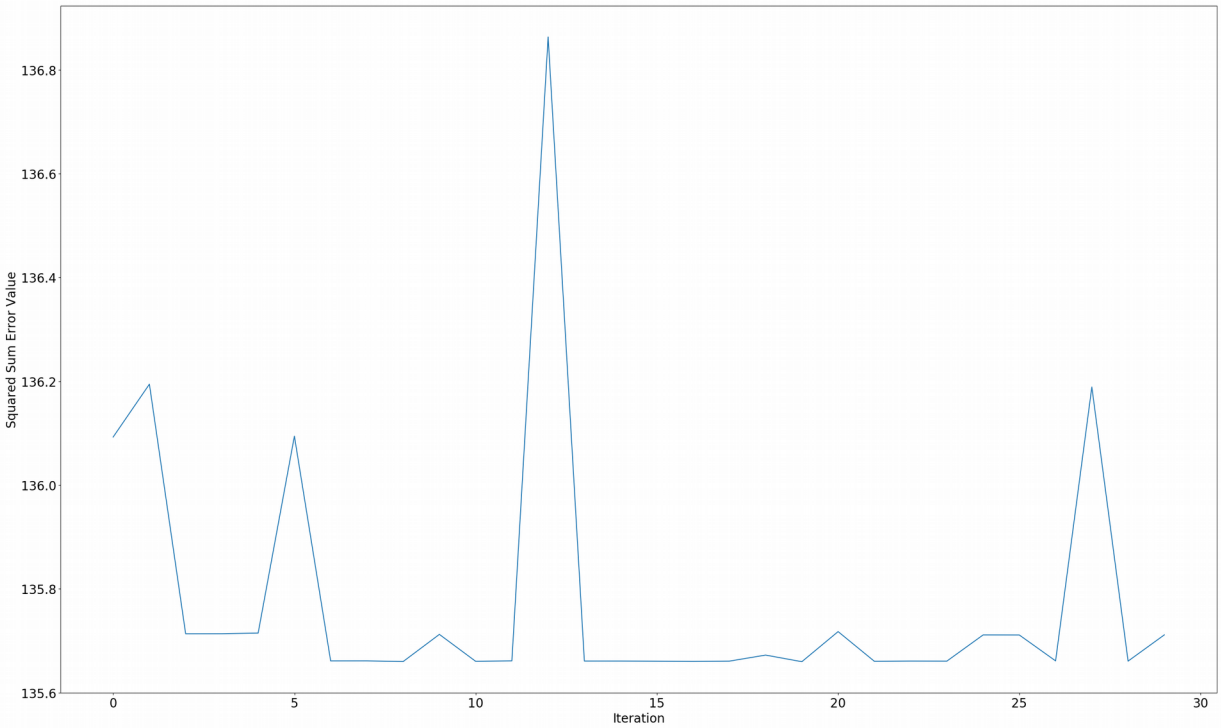
Final SSE value when running with 8 clusters originally positioned at first 8 points of the data: 135.541445399

Task 1.3

Discussion

By performing clustering with different cluster centres, and by inspecting the results of the 30 runs of K-means clustering with random centres we can observe the following:

- SSE generally converges to a similar number
- There is little difference between the maximum and minimum SSE observed

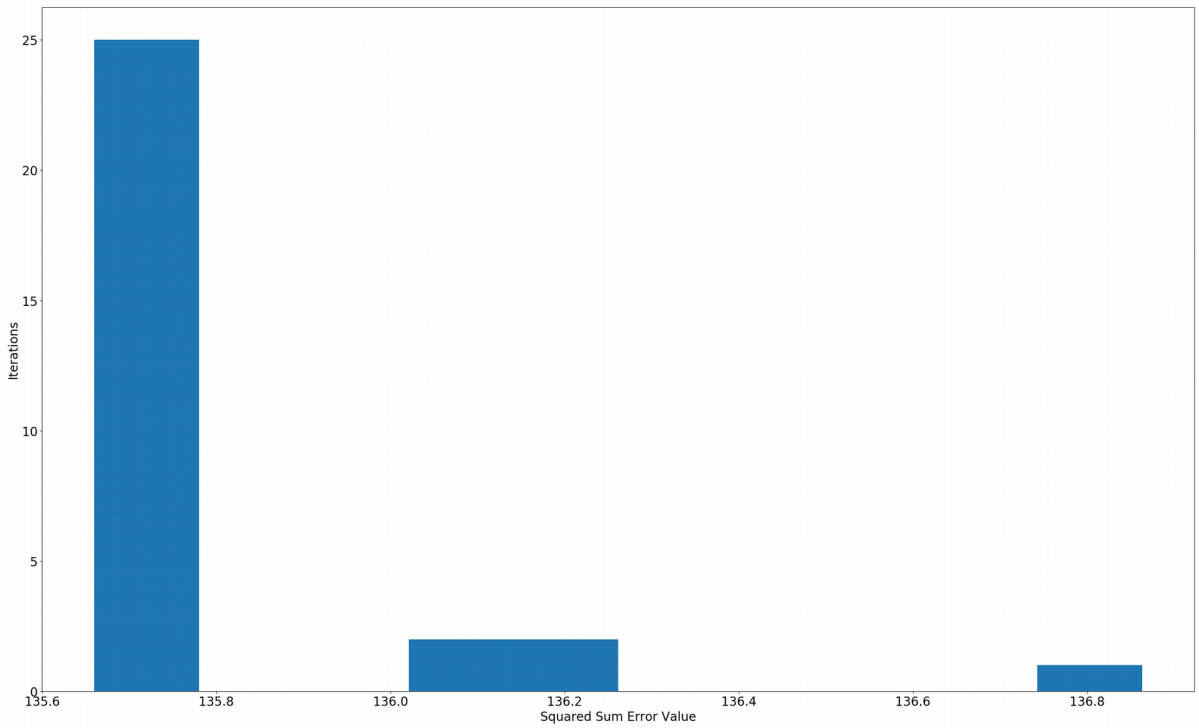


In my latest 30 runs of K-means clustering I have gathered the following data

Maximum SSE: 136.863054832

Minimum SSE: 135.659783905

Difference between minimum and maximum SSE: 1.20327092693



The plots above clearly reflect the convergence of Squared Sum Error to a single value.

Improving the performance of clustering

There are few ways in which we can improve our clustering performance:

- Better way to pick initial K centroids – it has been studied that by maximizing inter-cluster distances we get better clustering. So one way of improving the current clustering is choosing clusters so that we maximize the distance between them, instead of choosing them at random. However, we need to bare in mind that, as we saw in the 30 run experiment, different initial centres improve the performance just slightly.

- Number K of centroids – in the coursework we use 8 clusters. We can pick K value after we have studied the data and chosen a value that we seem fit. Another approach is to pick K randomly and dynamically converge or split clusters as the clustering runs on the basis of some observation.

- Setting a maximum limit on the number of times the algorithm runs - we currently stop running the clustering converges or reaches maximum iterations. However, in some cases centroids change only slightly after certain amount of runs or oscillate between values indefinitely. We currently have maximum amount of iterations, however, we can additionally measure the value of change in cluster position between runs and set a threshold, which once reached stops the clustering.

- Distance measure – we use Euclidean distance, which in some cases is not optimal. We can pick another distance measure or even a combination of such.