Domas Budrys – Assignment 7 CSCI5080

Question 1 (Time spent: 1h 45min):

a)

New centers:

1. $A_1(2, 10)$
2. $B_1(6, 6)$
3. $C_1(1.5, 3.5)$

b)

Final Clusters:

1. A_1, C_2, B_1
2. A_3, B_2, B_3
3. C_1, A_2

Question 2 (Time spent: 1h 30min):

Clustering-based support vector machines (CB-SVM) were created in order to improve and overcome issues of original SVMs. The main problem of SVMs are their limitation for large-scale data mining as for "pattern recognition or machine learning because the training complexity of SVMs is highly dependent on the size of a data set." To overcome such a problem, CB-SVM was designed explicitly to handle large data set containing from million to billions of data entries. This was achieved by applying a hierarchal micro-clustering algorithm "that scans the entire data set only once to provide an SVM with high quality samples that carry the statistical summaries of the data such that the summaries maximize the benefit of learning the SVM." Due to this, CB-SVM is easily scalable regarding to training efficiency while providing the highest performance of SVMs.

Experiments which were performed using CB-SVM on synthetic and real data sets have shown that it is highly scalable for large data sets and provides high classification accuracy.

Question 3 (Time spent: 2h):

********See completed projects in the submission folder.