Homework 2 (100 points). Due: Monday, Oct. 27, any time

Problem 1 (50 points)

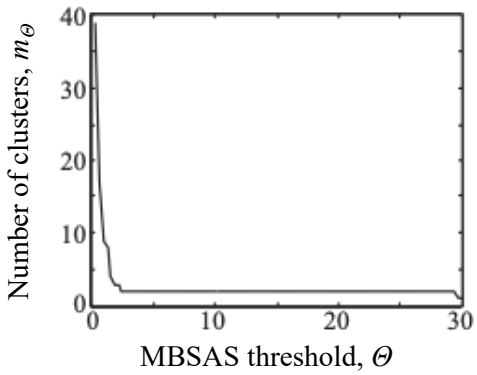
(a) (**15 points**) Implement your own MBSAS algorithm from the lecture slides.

(b) (**15 points**) Implement the estimator of the number of clusters (see Slide 17 of Unit 3. Part 2). Ensure that your implementation can also plot the dependency of the number of clusters, m_Θ, vs. the threshold, Θ.

Implementation notes:

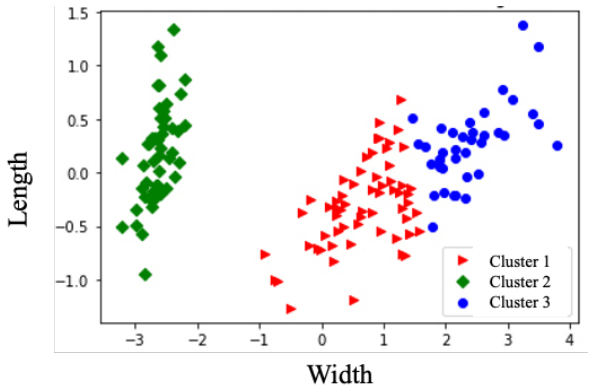
1) as discussed in the lecture, the optimal number of clusters m_Θ corresponds to the widest flat region in a plot that measures the dependence of m_Θ on the value Θ of the threshold in MBSAS.
2) If a segment $[\Theta_1, \Theta_2]$ represents such widest region, then any threshold value taken from $[\Theta_1, \Theta_2]$ will work. For instance, a standard way is to assign such Θ is: $\Theta = \frac{\theta_1+\theta_2}{2}$

An example of how such kind of graph would look like is:



(c) (**20 points**) Apply your MBSAS algorithm and the cluster number estimator to the real-world dataset stored in the attached file **cluster_data.txt**. **First**, run your number of cluster estimator and report: 1) the optimal number of clusters, m_Θ and 2) threshold Θ that gives you that number. **Second**, plot the X, Y coordinates for the entire dataset and their cluster assignments using the obtained threshold Θ. Use different symbols and colors to represent your data points for different clusters. Label X and Y axis as 'Width' and 'Length', correspondingly. Label each cluster as “Cluster 1”, “Cluster 2”, etc. **Last**, report the time it took to cluster this dataset.

An example of how such kind of plot could look like is:



Problem 2 (50 points)

(a) (**20 points**) Implement your own k-means algorithm from the lecture slides.

(b) (**15 points**) Using the k-means algorithm, cluster the data from the same file **cluster_data.txt.** Implementation notes:

- While k for our k-means is not known, we can use the estimates for m_Θ from Problem 1 , subproblem (c). And the for the k-means, assume k = m_Θ
- Plot the clustering result in the same way you plot the results for MBSAS clustering. The only difference that you should potentially see is the different cluster (that is, color) assignment for some data points.
- Report the time it took run the algorithm and compare with the time from Problem 1 (c). Explain the difference.

(c) (**15 points**) Investigate, what will happen if you "force" your k-means to use the incorrect number of clusters. For that, run your k-means algorithm for: 1) k = m_Θ + 1, and 2) k = m_Θ + 5. Plot the two new clusterings as two different plots (they will look similar to (b), but with more colors due to the greater number of clusters). Compare the two plots with the plot in (b) and draw conclusions.