ML Problem Solving

K-Means Clustering

1. Consider the following dataset, where each row is an observation. The first columns are input features.

$D = [[3, 3], [1, 1], [-1, 0], [2, 2], [-2, 2]]$

Draw the data points in a coordinate system to represent your input space.

2. Apply the K-means algorithm for the provided K number of clusters and using the following initializations:

- k=2, initial centroids: [3,3],[0,0]
- k=2, initial centroids: [2,2],[-2,-2]

3. Calculate the simplified silhouette $\sum_i \frac{b'_i - a'_i}{max(a'_i, 'b_i)}$, where b'_i is the average distance to other-cluster centroids and a'_i is the distance from the cluster centroid. What is the best case?

4. Calculate WSS (within-cluster sum of squares). What is the best case?

5. Now consider the following labeled version of the same set of datapoints (3rd column carries the label).

$D = [[3, 3, 1], [1, 1, 1], [-1, 0, 1], [2, 2, 0], [-2, 2, 0]]$

What is the accuracy for each previously calculated clustering result if the clusters are considered as classes? How the accuracy relates to the silhouette and the WSS?