K-Means Clustering

Q1. Briefly describe how the K-means clustering algorithm works.

[Answer]

In general, K-means is a heuristic algorithm that partitions a data set into *K* clusters by minimising the sum of squared distance in each cluster. The algorithm consists of three main steps: a) initialisation by setting seeding points (or initial centroids) with a given *K*, b) dividing all data points into *K* clusters based on *K* current centroids, and c) updating *K* centroids based on newly formed clusters. It can be shown that the algorithm always converges after several iterations of repeating steps b) and c).

Q2. Summarise the strength and the weakness of K-means clustering.

[Answer]

The strength of K-means algorithm lies in its computational efficiency and the nature of easy-to-use. In contrast, there are a number of weaknesses: a) requiring the prior knowledge of cluster numbers, *K*, b) sensitive to initialisation, which leads to unwanted solutions, c) sensitive to outliers and noise, which results in an inaccurate partition, d) incapable of handling clusters of a non-convex shape, and e) inapplicable to categorical data.

Q3. You are to cluster eight points: x_1 = (2, 10), x_2 = (2, 5), x_3 = (8, 4), x_4 = (5, 8), x_5 = (7, 5), x_6 = (6, 4), x_7 = (1, 2) and x_8 = (4, 9). Suppose, you assigned x_1, x_4 and x_7 as initial cluster centres for K-means clustering (k = 3). Using K-means with the Manhattan distance, compute the three clusters for each round of the algorithm until convergence.

[Answer]

In order to use the K-means algorithm, we need to calculate the distance between a point to a centroid. At the first round, we use the Manhattan distance measure to have

$$d_{21}=5,\ d_{31}=12,\ d_{51}=10,\ d_{61}=10,\ d_{81}=3;$$
$$d_{24}=6,\ d_{34}=7,\ \ d_{54}=5,\ \ d_{64}=5,\ \ d_{84}=2;$$
$$d_{27}=4,\ d_{37}=9,\ \ d_{57}=9,\ \ d_{67}=7,\ \ d_{87}=10;$$

After the first round, three clusters are { x_1}, { x_3, x_4, x_5, x_6, x_8}, and { x_2, x_7}.

With the grouping, we update three centroids to be (2, 10), ((8+5+7+6+4)/5, (4+8+5+4+9)/5) and ((2+1)/2, (2+5)/2); i.e., A: (2, 10), B:(6, 6), C: (1.5, 3.5).

Based on the new centroids, we recalculate the distances as follows:

$$d_{1A}=0,\ d_{2A}=5, d_{3A}=12, d_{4A}=5, d_{5A}=10,\ d_{6A}=10,\ d_{7A}=9,\ d_{8A}=3$$
$$d_{1B}=8,\ d_{2B}=5,\ d_{3B}=4,\ \ d_{4B}=3,\ d_{5B}=3,\ \ \ d_{6B}=2,\ \ \ d_{7B}=9,\ d_{8B}=5$$
$$d_{1C}=7,\ d_{2C}=2,\ d_{3C}=7,\ \ d_{4C}=8,\ d_{5C}=7,\ \ \ d_{6C}=5,\ \ \ d_{7C}=2,\ d_{8C}=8$$

After the second round, three clusters are { x_1, x_8}, { x_3, x_4, x_5, x_6}, and { x_2, x_7}.

At the third round, three centroids are A: (3, 9.5), B: (6.5, 5.25), (1.5, 3.5).

After the third round, three clusters are { x_1, x_8}, { x_3, x_4, x_5, x_6}, and { x_2, x_7}.

Since there is no change in three clusters in the second and the third round, the algorithm converges and the result after third round would be the final grouping result.