Sheet 6

Group 26 . Tu. 16:00
Ming Qu
Liuhan Pan

Assignment 2

$$\mu[C_i, \pi_{C_i}] = \sum_i d^2(C_{\pi_{C_i}}, P_i)$$

a)

Assign each data point to the closest center.

Suppose after the assignment, the updated clusters are $\pi_1^*, \pi_2^*, \dots \pi_k^*$. Then,

$$\mu(C_1, C_2, \dots C_k, \pi_1^*, \pi_2^*, \dots \pi_k^*) \le \mu(C_1, C_2, \dots C_k, \pi_1, \pi_2, \dots \pi_k)$$

Each center finds the centroid of points it owns:

$$C_i = \arg\min_c \sum_{P_j \in \pi_i} d(P_j, c)$$

For Euclidean distance measure: $C_i = \frac{1}{|\pi_i|} \sum_{P_j \in \pi_i} P_j$

After the update:

$$\mu(C_1^*, C_2^*, \dots C_k^*, \pi_1, \pi_2, \dots \pi_k) \le \mu(C_1, C_2, \dots C_k, \pi_1, \pi_2, \dots \pi_k)$$

Thus, we can see that:

when holding P_i fixed, μ decreases with respect to C_{π_i}

when holding C_{π_i} fixed, μ decreases with respect to P_i.

So, the error function μ strictly decreases in each step.

b)

There are only a finite number of ways of partitioning X data into k groups. So, there are only a finite number of points to cluster assignments.

c) As showed in b), there are a finite number of cluster assignments, if the algorithm tried to run forever, ~~there must~~ it must end up passing through a given assignment more than once. But the k-means algorithm strictly reduce the error on each step (as showed in (a)), so the algorithm cannot come back to the same assignment.

Assignment 3

Discussing three cases:

① If choosing initial parameters appropriately, for example the initial paras close to the real means, then the iteration times will decrease, and the computational times will decrease, result will be accurate.

② If choosing the initial values very close to each other (or some values overlay) usually cause them to compete for the same cluster, which will cause to reach local maxima very quickly, the result won't be accurate. Computational time mainly depends on iteration times and how many Gaussian functions overlay. So the computational time will decrease.

③ If initial values won't compete for the same cluster, but very far from the real means. Then, reaching local maximum will be slow, iteration time will increase, computational time will increase. Result will be accurate.