Multimedia oral examination - Unsupervised Learning

1.What is clustering

Given a dataset with N items,cluster algorithm make a partition on the dataset into k groups， k is a hyperparameter given by user.

2.K-means

(1)the general procedure of k-means

K-means is a general solution for clustering.

Given N items and the number of cluster amount K, K-means select K items as initial centers,and **iteratively assign** items to its closest center and update each center with **average** of items in this group,until centers do not change.

(2)pros and cons

It is simple and fast.The complexity is O(K·N·D),which is the efficient. It only has one parameter and can converge quickly(within 20 iterations). The clustering quality are moderately good in most of the cases

But K is given by user and should be carefully tuned. It only obtains sub-optimal solution, this is true for all clustering algorithms. It is slow in high dimension and big data size.

3.K-means++

The motivation of K-means++ is to **optimize the initialization** of clustering centers by **selecting points far apart from each other**.

The advantage is that K-means++ leads to faster convergence and **better adaptation** to the data distribution.

Initialization procedure is modified by selecting one item randomly as the first center and repeat center selecting K-1 times. Rest of initial centers are selected by calculate distance for each item to existing centers.The probability of being a new center for each item is the square of its closest distance.

4.Boost k-means(k-means#)

k-means is slow in high dimension and big data size with complexity of O(n·d·k·t). k-means# reduce k to log(k) by **hierarchical** clustering and make t smaller by faster converges.

k-means is formulated as minimizing the sum of distance between each point and its center.This minimization problem can be transformed to a maximization problem. To maximize this optimization function,the procedure is pick each item randomly and move it to another cluster if this opeartion will increase the optimization function.

In k-means#,it is unnecessary to do **initial assignment** and **seeking closest centroid**.