1. **Clustering problem**

Clustering is to divide data into corresponding classes according to the similarity between data. The similarity between the same class is very high, while the difference between different classes is the greatest.

1. **Hierarchical clustering**

Hierarchical clustering, as the name suggests, is to cluster layer by layer. Firstly, the distance between samples is calculated, and the nearest point is merged into the same class each time. Then, the distance between classes is calculated, and the nearest class is merged into a large class. Keep merging until a class is composed. The calculation methods of the distance between classes are: the shortest distance method, the longest distance method, the middle distance method, the class average method and so on. According to the order of hierarchical decomposition, hierarchical clustering algorithm is divided into cohesive hierarchical clustering algorithm and split hierarchical clustering algorithm. Agglomerative hierarchical clustering starts from the point as an individual cluster and merges two closest clusters at each step, and finally forms a tree graph.

The steps of basic agglomerative hierarchical clustering algorithm are as follows:

1: If necessary, calculate the proximity matrix

2: repeat

3: Merge the two closest clusters

4: The proximity matrix is updated to reflect the proximity between the new cluster and the original cluster

5: Until there is only one cluster left / a termination condition reached

1. **The differences between hierarchical clustering and k-means**

The basic process of hierarchical clustering is as follows: each sample point is regarded as a cluster; the distance between each cluster is calculated, and the nearest two clusters are clustered into a new cluster; the above process is repeated until there is only one cluster.

The process of K-means is as follows: firstly, take any k sample points as the initial centers of K clusters; for each sample point, calculate the distance between them and K centers, and classify it into the cluster with the smallest distance center; when all the sample points are classified, recalculate the centers of K clusters; repeat the above process until the cluster that the sample points belong to no longer changes.

The time complexity of K-means is low, but it needs to specify the number of clusters before clustering, that is, K value. When the amount of data is large, it is difficult to judge in advance.

Hierarchical clustering does not need to set the number of clusters in advance, only focus on the distance between clusters, and eventually form a tree graph, which can find the hierarchical relationship of clusters, but the computational complexity is high, and the algorithm is likely to cluster into a chain.