K-Means Algorithm

First part:

K-means algorithm is an unsupervised clustering algorithm, which tries to partition the dataset into K pre defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group, using iterative method. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible.

K-means算法是一种无监督聚类算法，它尝试将数据集划分为K个预先定义的不同的不重叠子群(聚类)，其中每个数据点只属于一个组，采用迭代的方法。它试图使集群内的数据点尽可能相似，同时使集群尽可能不同(远)。

K-Means algorithm is an iterative algorithm that tries to partition the dataset into K pre defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster’s centroid (arithmetic mean of all the data points that belong to that cluster) is at the minimum. The less variation we have within clusters, the more homogeneous (similar) the data points are within the same cluster.

K- means算法是一种迭代算法，它试图将数据集划分为K个预先定义的不同的不重叠的子组(簇)，其中每个数据点只属于一个组。它试图使集群内的数据点尽可能相似，同时使集群尽可能不同(远)。它将数据点分配给一个集群，使数据点与集群质心(属于该集群的所有数据点的算术平均值)之间的平方距离之和最小。集群内的变化越少，同一集群内的数据点就越相似。

Principle(central idea):

pre-determined constant K, constant K means that the final number of cluster categories, first randomly selected initial point as the center of mass, and by calculating the similarity between each sample and center of mass (here as the Euclidean distance), similar to the sample points to the classes, and then to recalculate the center of mass of each class (that is, for the class center), repeat this process, until the center of mass is no longer change, eventually will determine the category of each sample and the center of mass of each class.

原则(中心思想):

预先确定的常数K,常数K意味着集群的最后数量分类,首先随机选择初始点为中心的质量,通过计算每个样本之间的相似性和质心(这里欧几里得距离),类似于采样点的类,然后重新计算每个类的质心(即类中心),重复这个过程,直到重心不再变化,最终将确定每个样本的类别和每个类别的质心。

Algorithm Flow:

1. Select the number of clustering K (when kmeans algorithm transfers hyperparameters, it only needs to set the maximum k value)

2. Randomly generate K clustering, and then determine the clustering center, or directly generate K centers.

3. Determine the clustering center of each point.

4. Then calculate the new clustering center.

5. Repeat the above steps until the convergence requirements are met.(This is usually the fixed center point that no longer changes.)

算法流程:

1. 选择聚类个数K(当kmeans算法传输超参数时，只需要设置K的最大值)

2. 随机生成K个聚类，然后确定聚类中心，或者直接生成K个中心。

3.确定每个点的聚类中心。

4. 然后计算新的聚类中心。

5. 重复上述步骤，直到满足收敛要求。(这通常是不再改变的固定中心点。)

The way K-Means algorithm works is as follows:

1. Specify number of clusters K.

2. initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.

3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn’t changing.

• Compute the sum of the squared distance between data points and all

centroids.

• Assign each data point to the closest cluster (centroid).

• Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

K-Means算法的工作方式如下:

1. 指定集群数量K。

2. 初始化中心首先洗牌数据集，然后为中心随机选择K数据点不替换。

3.继续迭代，直到中心没有变化。也就是说，数据点到集群的分配没有改变。

•计算数据点与所有数据点之间距离的平方之和

重心。

•将每个数据点分配到最近的簇(质心)。

•通过取属于每个集群的所有数据点的平均值来计算集群的中心。

Advantages:

1. Simple principle (near the center point), easy to implement

2. Upper in clustering effect (depending on the choice of K)

3. Space complexity O (N) Time complexity O (IKN)

N is the number of sample points, K is the number of center points, and I is the number of iterations

Disadvantages:

1. Noise sensitive to outliers (the center point is easily shifted)

2. It is difficult to find clusters with very different sizes and carry out incremental calculation

3. The result is not necessarily the global optimal, but only the local optimal (related to the number of K and the selection of initial value)

优点：

1、原理简单（靠近中心点） ，实现容易

2、聚类效果中上（依赖K的选择）

3、空间复杂度o(N)时间复杂度o(IKN)

N为样本点个数，K为中心点个数，I为迭代次数

缺点：

1、对离群点， 噪声敏感 （中心点易偏移）

2、很难发现大小差别很大的簇及进行增量计算

3、结果不一定是全局最优，只能保证局部最优（与K的个数及初值选取有关）