KMeans聚类 K值的确定以及初始类簇中心点的选取

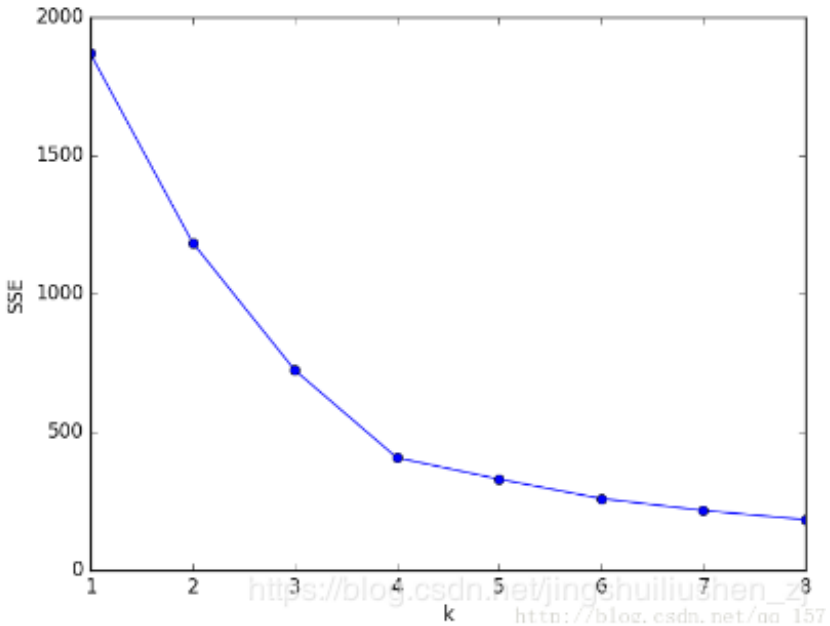
blog.csdn.net/jingshuiliushen_zj/article/details/83754306

KMeans算法是最常用的聚类算法，基本思想是:在给定K值和K个初始类簇中心点的情况下，把每个样本点分到离其最近的簇中，然后重新计算每个簇的中心点(取平均值)，然后再迭代的进行分配点和更新类簇中心点的步骤，直至类簇中心点的变化很小，或者达到指定的迭代次数。

KMeans算法本身思想比较简单，但是确定一个合适的K值和K个初始类簇中心点对于聚类效果的好坏有很大的影响。

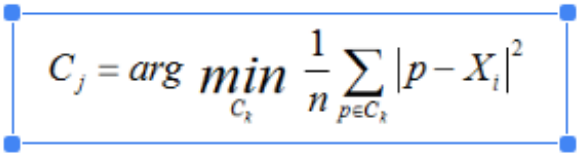
K值的确定

1、样本聚类误差平方和，核心指标是SSE(sum of the squared errors，误差平方和)
SSE=∑Kk=1∑p∈Ck|p−mk|2SSE=∑k=1K∑p∈Ck|p−mk|2 SSE=\sum_{k=1}^K\sum_{p \in C_k}|p-m_k|^2SSE=k=1∑K p∈Ck ∑ |p−mk |2
其中，K是聚类数量，p是样本，mkmk m_kmk 是第k个聚类的中心点。K越大，SSE越小，说明样本聚合程度越高。
当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，这个最先趋于平缓的点就是合适的K值。



2、轮廓系数法
某个样本点XiXi X_iXi 的轮廓系数定义如下：

S=b−amax(a,b)S=b−amax(a,b) S=\frac{b-a}{max(a,b)}S=max(a,b)b−a
其中，a是XiXi X_iXi 与同簇的其他样本的平均距离，称为凝聚度，b是XiXi X_iXi 与最近簇中所有样本的平均距离，称为分离度。而最近簇的定义是



其中p是某个簇CkCk C_kCk 中的样本。就是用XiXi X_iXi 到某个簇所有样本平均距离作为衡量该点到该簇的距离后，选择离XiXi X_iXi 最近的一个簇作为最近簇。
求出所有样本的轮廓系数后再求平均值就得到了平均轮廓系数。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。

初始类簇中心点的确定

1、选择批次距离尽可能远的K个点
首先随机选择一个点作为第一个初始类簇中心点，然后选择距离该点最远的那个点作为第二个初始类簇中心点，然后再选择距离前两个点的最近距离最大的点作为第三个初始类簇的中心点，以此类推，直至选出K个初始类簇中心点。
2、 选用层次聚类或者Canopy算法进行初始聚类，然后利用这些类簇的中心点作为KMeans算法初始类簇中心点。