2. K-means

1.原理

1.1 算法过程

1.初始化k个质心，作为初始的k个簇的中心点，k为人工设定的超参数；

2.然后对于每一个样本分别计算其k个质心的距离，并将样本点归于最近的一类中。

3.重新计算质心，即将每一类中的所有点取平均值。

4.重复上述过程直到达到预定的迭代次数或质心不再发生明显变化

1.2损失函数

$$SSE = \sum_{k=1}^{K} \sum_{p \in C_k} |p - m_k|^2$$

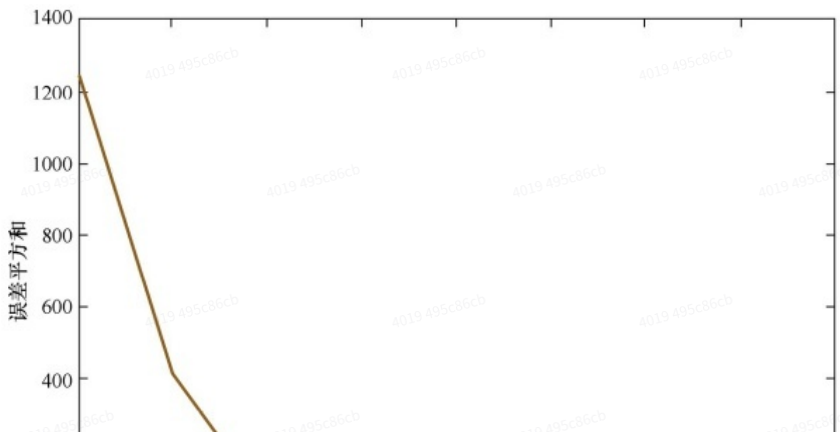
其中，K是聚类数量，p是样本，mk是第k个聚类的中心点。SSE越小，说明样本聚合程度越高。

1.3怎么确定聚类数量K（聚类如果不清楚有多少类，有什么方法？）

和评估分类或回归的方式一样，选择某个metric或某些metrics下最好的k，例如sse（其实就是kmeans的损失函数了），轮廓系数.

k的大小调参，手工方法，手肘法为代表。

手肘法其实没什么特别的，纵轴是聚类效果的评估指标，根据具体的问题而定，如果聚类是作为单独的任务存在则使用sse或轮廓系数这类无监督的metric作为纵坐标，然后找到metric最好并且k最小的结果对应的k为最终的选择；（我们说的学习曲线）



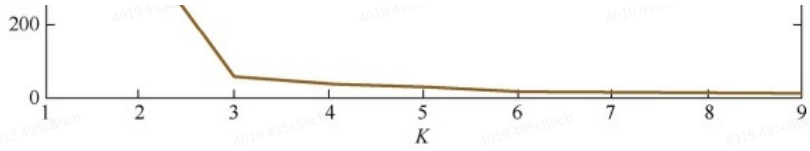


图5.3　K均值算法中K值的选取：手肘法

1.4k-means的缺点，怎么解决?

1. 对异常样本很敏感，簇心会因为异常样本被拉得很远

解决方法即做好预处理，将异常样本剔除或修正

2. k值需要事先指定，有时候难以确定

解决方法即针对k调参。

3. 只能拟合球形簇

对于流形簇等不规则的簇或是存在簇重叠问题的复杂情况等，效果较差。

解决方法，换算法。

4. 无法处理离散特征，缺失特征