面试常考的算法知识点总结 —— kmeans

原创 稀饭的写作小屋 稀饭居然不在家 2月27日

收录于话题
#数据分析算法总结

24个

> 作者：稀饭

本文约**1400字**，建议阅读**13分钟**。

1、什么是kmeans算法？

答：kmeans是一种动态聚类的方法。基本的kmeans算法的思想很简单，事先确定常数k，常数k意味着最终的聚类类别数，首先随机选定初始点为质心，并通过计算每一个样本与质心之间的相似度(这里为欧式距离)，将样本点归到最相似的类中，接着，重新计算每个类的质心(即为类中心)，重复这样的过程，知道质心不再改变，最终就确定了每个样本所属的类别以及每个类的质心。由于每次都要计算所有的样本与每一个质心之间的相似度，故在大规模的数据集上，kmeans算法的收敛速度比较慢。

2、kmeans算法有什么优缺点？

答：（1）优点

[1] 算法简单、迅速；

[2] 对于处理大数据集，该算法是相对可伸缩和高效的，因为它的复杂度大约是O (nkt)。其中n是所有对象的数目，k是分类的数目，t是迭代的次数，该算法经常以局部最优结束；

[3] 当类是密集、球状或者团状，且类与类之间区别明显时，该算法聚类效果很好。

（2）缺点

[1] 该算法只有在类的平均值被定义的情况下才能使用，不适用于某些分类属性的数据；

[2] 对初值比较敏感，对于不同的初始值可能会导致不同的聚类结果；

[3] 不适合于发现非凸面形状的类，或者大小差别很大的类；

[4] 对于“噪声”和孤立点数据敏感，少量的该类数据能够对平均值产生极大影响。

3、在使用kmeans算法的时候需要注意哪些问题？

答：（1）算法中的k值需要认真选取；

（2）要慎重选取初始的聚类中心，如果选择不当可能很容易陷入局部最优；

（3）样本要随机选取，可以提高算法的收敛速度。

4、简述kmeans算法的基本步骤？

答：（1）**第一步**：判断样本集可以分为几类，设定好类个数k；

（2）**第二步**：在样本集X中，随机选择k个数据点作为初始聚类的中心；

（3）**第三步**：计算样本集中每一个数据点到这k个聚类中心的距离，一共nk个距离；

（4）**第四步**：将每个数据点归到离它最近的聚类中心的类别中，重复n次，直到每一个数据点都进行了归类（对于已经设定为类中心的点，其到它自己的距离最小，为0）；

（5）**第五步**：待所有样本点归类完成后，重新计算每一类的中心，并计算误差衡量指标；

（6）**第六步**：比较误差衡量指标是否在给定阈值内，如果小于等于阈值，输出分类结果；如果大于阈值，以新得到的聚类中心，重复“第三步 → 第五步”，直到收敛。

5、如何确定kmeans算法中的k值？

答：可以采用轮廓系数法。在实际应用中，由于kmeans一般作为数据预处理，或者用于辅助分聚类贴标签。所以k一般不会设置很大。可以通过枚举，令k从2到一个固定值如10，在每个k值上重复运行数次kmeans(避免局部最优解)，并计算当前k的平均轮廓系数，最后选取轮廓系数最大的值对应的k作为最终的集群数目。

6、分析异常点和初值对kmeans算法的影响？

答：（1）**异常点**：kmeans算法在迭代的过程中使用所有点的均值作为新的质点(中心点)，如果簇中存在异常点，将导致均值偏差比较严重。比如一个簇中有2、4、6、8、100五个数据，那么新的质点为24，显然这个质点离绝大多数点都比较远；在当前情况下，使用中位数6可能比使用均值的想法更好，使用中位数的聚类方式叫做k-mediods聚类(k中值聚类)。

（2）**初值**：kmeans算法是初值敏感的，选择不同的初始值可能导致不同的簇划分规则。为了避免这种敏感性导致的最终结果异常性，可以采用初始化多套初始节点构造不同的分类规则，然后选择最优的构造规则。针对这点后面因此衍生了：二分kmeans算法、kmeans++算法、kmeans||算法、canopy算法等。