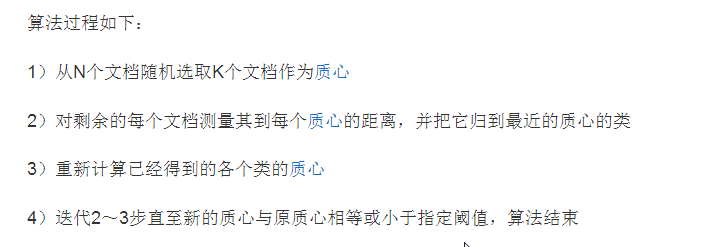
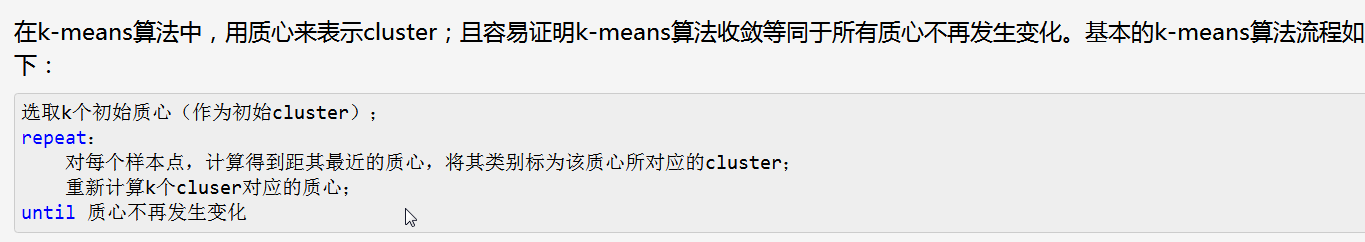
简介：

<https://baike.baidu.com/item/K-means/4934806>



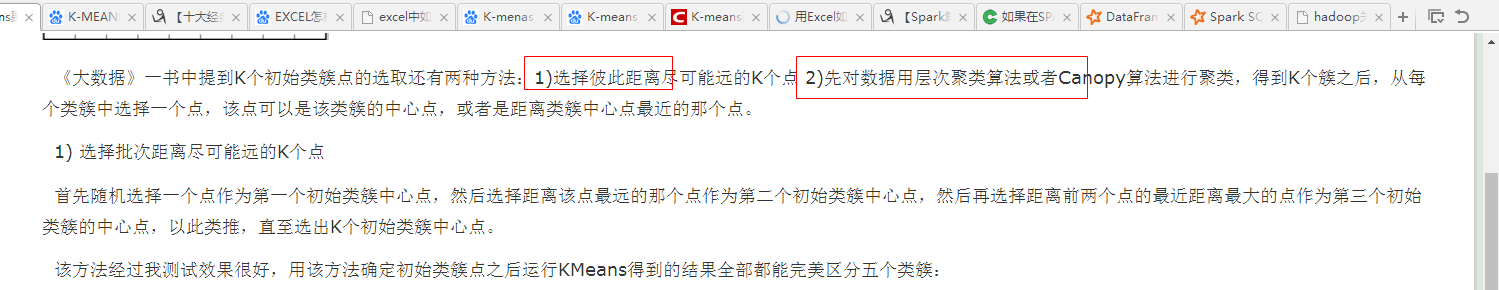
计算流程如下，以此为准：



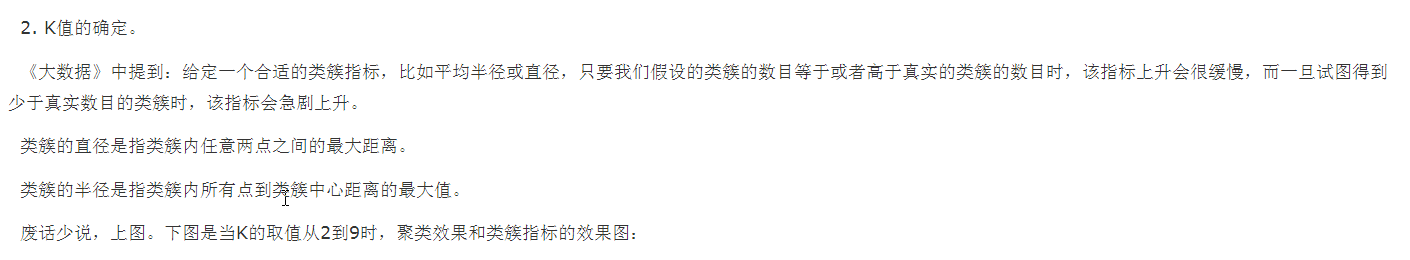
1.如何确定K的值，以及初始K个质心的选择？

见：<https://www.cnblogs.com/zhizhan/p/4816884.html>

质心选择

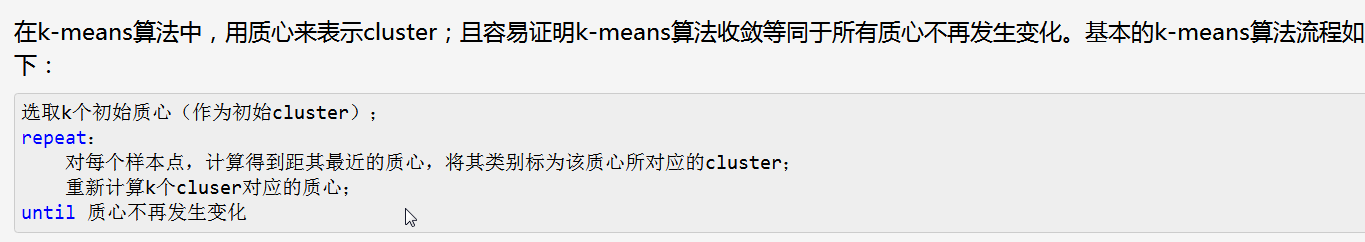


K值选择：



1. 相似度：属于同一类簇的点，相似度较高。 使用：质心的概念，常用的是点之间的“欧式距离”————对于多维向量，则是该向量的范数,即

\*\*：相似度，用于计算在每一次迭代中，某点属于哪一类。



1. 对每个点计算距其最近的质心：

步骤0：初始确定了k和k个质心，

步骤1：For 每个样本点i：

For 每个质心j：

Distance（ij） = 欧拉距离（i，j）

MinDistance(i,m) = Min(Distance(i,j)) #m是第m个质心，点i离m欧式距离最

点i属于 第m类 #确定新的k个类簇

步骤2：compute(第二组质心){

For 每个新的类簇t

count\_n # 第n类的样本点个数

Centor\_point = sum(Vector(i))/count\_n #一个点就是1个向量，把一个类的所有向量求平均，就得到一个“质心”

}

迭代1次：

重复步骤1——即（求各点到各质心举例，选最小举例，再归类）

重复步骤2——即（重新计算质心）

迭代2次：

重复步骤1

重复步骤2

迭代3次：

重复步骤1

重复步骤2

...

迭代n次：

重复步骤1

重复步骤2

直至质心不再发生变化， 则停止迭代。 最终确定K值，及K个类。

当然，相似度可能可以用其他指标度量。

CART-分类算法：

<https://blog.csdn.net/baimafujinji/article/details/53269040>

我目前的理解——CART生成决策树步骤：

1. 确定各属性在树中的顺序（第1层，第二层...）
2. 确定每个属性的“分类规则”————二元、多元离散、连续

当然，生成决策树后，还可以根据剪枝，（去除不明确的分类）。

一个例子，想找出 “拖欠贷款者”——具有哪些特征——————符合该特征分类路径的，则认定为“拖欠贷款者”。

对于多元离散属性，如何选择分类方式？

假设“婚姻与否”属性，包含多个值 （M，D，S） ，M-Married ,D-Divoriced,S-Single

可以分为3组：{M | D,S} 、{ D | M,S} 、{S | M,D}

（肯定是分成上面3组，如果是K元属性，则分为K组，分组规则——该属性的某1个值为1类，其他所有值为另1类，完全按照2元划分）

如何确定哪一组作为该属性的“分类方法”（即度量子节点的不纯度，不纯度越高，说明分类越不明确，则越不应该作为分类规则）————Gini系数。（根据该组的属性值、因变量的属性值，计算4种情况出现的概率，并据此计算Gini系数）

找出Min(Gini)最小的这一组，作为该属性的“分类规则”

CART分类算法，最后会——构造出一条决策树。

如果一条记录（多个属性值，组成的向量），符合这个决策树的某条路径，则可得知对应的“因变量值”。 因变量必然是0-1分类吧。

对于多元连续属性，如何选择分类方式?

先对k个连续数值，升序排序，求出k-1个中间值（即相邻两项的平均数），然后分别以这k-1个”界限”， 求Gini系数（计算不纯度）————就是（<=界限，则为1类;>界限，则为另1类）

如何确定 n个属性的分类顺序呢。

貌似要计算各个属性对因变量属性划分的信息增益（也是用Gini系数求）

“信息增益”————也就是不确定性的减少程度（熵减）

信息增益值，按降序排序，最大的作为最先分类。

还涉及到“剪枝”，什么时候“分类”停止的规则。