**算法关键点说明**

# 

郭娟

2018/3/8

# **一、 Kmeans算法逻辑（聚类算法，无监督**）

1. 任意选取K个点作为初始聚类中心
2. 计算每个样本到这K个聚类中心的距离，并将该样本与距离最小的聚类中心归为同一类。（距离可以是欧式距离，曼哈顿距离，余弦距离，切比雪夫距离等等）
3. 重新更新聚类中心，比如计算同一类别的质心作为新的聚类中心
4. 重复2， 3 直到新的聚类中心不发生变化，或聚类次数已经达到要求。

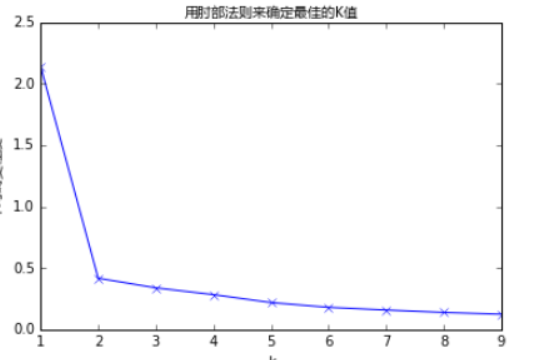
简单理解：

1. 首先输入k的值，即我们希望将数据集经过聚类得到k个分组。
2. 从数据集中随机选择k个数据点作为初始大哥（质心，Centroid）
3. 对集合中每一个小弟，计算与每一个大哥的距离（距离的含义后面会讲），离哪个大哥距离近，就跟定哪个大哥。
4. 这时每一个大哥手下都聚集了一票小弟，这时候召开人民代表大会，每一群选出新的大哥（其实是通过算法选出新的质心）。
5. 如果新大哥和老大哥之间的距离小于某一个设置的阈值（表示重新计算的质心的位置变化不大，趋于稳定，或者说收敛），可以认为我们进行的聚类已经达到期望的结果，算法终止。
6. 如果新大哥和老大哥距离变化很大，需要迭代3~5步骤。

# **二、肘部法则确定聚类个数**

通过肘部法则来估计聚类数量。肘部法则会把不同值的成本函数值画出来。随着值的增大，平均畸变程度会减小；每个类包含的样本数会减少，于是样本离其重心会更近。但是，随着值继续增大，平均畸变程度的改善效果会不断减低。 值增大过程中，畸变程度的改善效果下降幅度最大的位置对应的 值就是肘部。

成本函数：成本函数是各个类畸变程度（distortions）之和。每个类的畸变程度等于  
该类重心与其内部成员位置距离的平方和。



肘部

# **三、随机森林**

1. 用有抽样放回的方法（bootstrap）从样本集中选取n个样本作为一个训练集
2. 用抽样得到的样本集生成一棵决策树。在生成的每一个结点：
3. 随机不重复地选择d个特征
4. 利用这d个特征分别对样本集进行划分，找到最佳的划分特征（可用基尼系数、增益率或者信息增益判别）
5. 重复步骤1到步骤2共k次，k即为随机森林中决策树的个数。
6. 用训练得到的随机森林对测试样本进行预测，并用票选法决定预测的结果。

参考文献：

<http://blog.csdn.net/mao_xiao_feng/article/details/52728164>

<https://segmentfault.com/a/1190000007463203>

# **四、随机森林确定特征重要性**

计算某个特征X的重要性时，具体步骤如下：

1. 对每一颗决策树，选择相应的袋外数据（out of bag，OOB）​计算袋外数据误差，记为errOOB1.

所谓袋外数据是指，每次建立决策树时，通过重复抽样得到一个数据用于训练​决策树，这时还有大约1/3的数据没有被利用，没有参与决策树的建立。这部分数据可以用于对决策树的性能进行评估，计算模型的预测错误率，称为袋外数据误差。

1. ​随机对袋外数据OOB所有样本的特征X加入噪声干扰（可以随机改变样本在特征X处的值），再次计算袋外数据误差，记为errOOB2。
2. 假设森林中有N棵树，则特征X的重要性=∑（errOOB2-errOOB1）/N。这个数值之所以能够说明特征的重要性是因为，如果加入随机噪声后，袋外数据准确率大幅度下降（即errOOB2上升），说明这个特征对于样本的预测结果有很大影响，进而说明重要程度比较高。