K-Means算法

一、简介

K-means算法是硬聚类算法，是典型的基于原型的目标函数聚类方法的代表，它是数据点到原型的某种距离作为优化的目标函数，利用函数求极值的方法得到迭代运算的调整规则。K-means算法以欧式距离作为相似度测度，它是求对应某一初始聚类中心向量V最优分类，使得评价指标J最小。算法采用误差平方和准则函数作为聚类准则函数。

二、算法原理

在数据集中根据一定策略选择K个点作为每个簇的初始中心，然后观察剩余的数据，将数据划分到距离这K个点最近的簇中，也就是说将数据划分成K个簇完成一次划分，但形成的新簇并不一定是最好的划分，因此生成的新簇中，重新计算每个簇的中心点，然后在重新进行划分，直到每次划分的结果保持不变。在实际应用中往往经过很多次迭代仍然达不到每次划分结果保持不变，甚至因为数据的关系，根本就达不到这个终止条件，实际应用中往往采用变通的方法设置一个最大迭代次数，当达到最大迭代次数时，终止计算。

k-means的计算方法如下：

1、随机选取k个中心点；

2、遍历所有数据，将每个数据划分到最近的中心点中；

3、计算每个聚类的平均值，并作为新的中心点；

4、重复2-3，直到这k个中心点不再变化（收敛了），或执行了足够多的迭代；

时间复杂度：O(I\*n\*k\*m)

空间复杂度：O(n\*m)

其中m为每个元素字段个数，n为数据量，I为跌打个数。一般I,k,m均可认为是常量，所以时间和空间复杂度可以简化为O(n)，即线性的。

在实际应用中，由于K-means一般作为数据预处理，或者用于辅助分类贴标签。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k-means（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。

三、算法过程

1）从N个文档随机选取K个文档作为质心

2）对剩余的每个文档测量其到每个质心的距离，并把它归到最近的质心的类

3）重新计算已经得到的各个类的质心

4）迭代2～3步直至新的质心与原质心相等或小于指定阈值，算法结束

具体如下：

输入：k, data[n];

（1） 选择k个初始中心点，例如c[0]=data[0],…c[k-1]=data[k-1]；

（2） 对于data[0]….data[n]，分别与c[0]…c[k-1]比较，假定与c[i]差值最少，就标记为i；

（3） 对于所有标记为i点，重新计算c[i]={ 所有标记为i的data[j]之和}/标记为i的个数；

（4） 重复(2)(3)，直到所有c[i]值的变化小于给定阈值。

公式：



四、适用范围

K-means算法试图找到使平凡误差准则函数最小的簇。当潜在的簇形状是凸面的，簇与簇之间区别较明显，且簇大小相近时，其聚类结果较理想。前面提到，该算法时间复杂度为O(I\*n\*k\*m)，与样本数量线性相关，所以，对于处理大数据集合，该算法非常高效，且伸缩性较好。但该算法除了要事先确定簇数K和对初始聚类中心敏感外，经常以局部最优结束，同时对“噪声”和孤立点敏感，并且该方法不适于发现非凸面形状的簇或大小差别很大的簇。

五、优缺点

K-Means的主要优点有：

1）原理比较简单，实现也是很容易，收敛速度快。

2）聚类效果较优。

3）算法的可解释度比较强。

4）主要需要调参的参数仅仅是簇数k。

K-Means的主要缺点有：

1）K值的选取不好把握

2）对于不是凸的数据集比较难收敛

3）如果各隐含类别的数据不平衡，比如各隐含类别的数据量严重失衡，或者各隐含类别的方差不同，则聚类效果不佳。

4）采用迭代方法，得到的结果只是局部最优。

5）对噪音和异常点比较的敏感。

**bagging（装袋）方法**

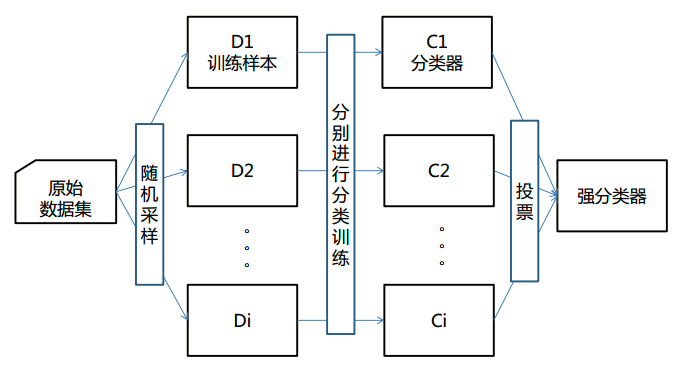
一、简介

① Bagging又叫自助聚集，是一种根据均匀概率分布从数据中重复抽样（有放回）的技术。

② 每个抽样生成的自助样本集上，训练一个基分类器；对训练过的分类器进行投票，将测试样本指派到得票最高的类中。

③ 每个自助样本集都和原数据一样大

④ 有放回抽样，一些样本可能在同一训练集中出现多次，一些可能被忽略。



由于Bagging算法每次都进行采样来训练模型，因此泛化能力很强，对于降低模型的方差很有作用。当然对于训练集的拟合程度就会差一些，也就是模型的偏倚会大一些。

二、基本思想

1.给定一个弱学习算法和一个训练集；

2.单个弱学习算法准确率不高；

3.将该学习算法使用多次,得出预测函数序列,进行投票；

4.最后结果准确率将得到提高。

三、算法流程

输入为样本集D={(x,y1),(x2,y2),...(xm,ym)}，弱学习器算法, 弱分类器迭代次数T。

输出为最终的强分类器f(x)

1）对于t=1,2...,T：

a)对训练集进行第t次随机采样，共采集m次，得到包含m个样本的采样集Dm

b)用采样集Dm训练第m个弱学习器Gm(x)

2）如果是分类算法预测，则T个弱学习器投出最多票数的类别或者类别之一为最终类别。如果是回归算法，T个弱学习器得到的回归结果进行算术平均得到的值为最终的模型输出。