KNN近邻算法

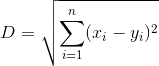
* 概念

存在一个训练样本集合A，在给定测试样本b时，基于某种距离度量，找出训练集A中与测试样本b最靠近的k个训练样本（k<20），之后基于这k个训练样本的信息来预测种类或值。

* 近邻算法的分类

在近邻算法中，根据处理的问题的不同，可以分为：

* 近邻分类算法：KNN用来预测种类。一般使用“投票法”，选择这k个样本中出现最多的类别来作为测试样本的类别。
* 近邻回归算法：KNN预测一个值。使用“平均法”，将k个样本的实值输出的平均值作为测试样本的输出。一般距离度量选择欧氏距离：

样本与样本之间的欧式距离

* KNN算法的流程
* 计算测试数据与各个训练数据之间的距离；
* 按照距离的递增关系进行排序；
* 选取距离最小的K个点；
* 确定前K个点所在类别的出现频率；
* 返回前K个点中出现频率最高的类别作为测试数据的预测分类。
* KNN算法的不足：
* 没有明显的训练过程，它是“懒惰学习”的典型代表，它在训练阶段所做的仅仅是将样本保存起来，如果训练集很大，必须使用大量的存储空间，训练时间开销为零。
* 在K-NN算法中，每一个预测样本需要与所有的训练样本计算相似度，计算量比较大。比较常用的方法有K-D树，局部敏感哈希等等
* 训练样本在训练集中所占权重是一样的，而实际可能有某个训练样本权重大或者权重小，这时应采用加权的方式。