## 3.KNN算法

### 3.1算法主要思想：

设训练数据集为 ，根据一定的距离度量规则，来寻找给定的数据的最相近的k个数据组成的集合，由集合来决定所属的类别。

即 ，

其中当时为1，否则为0

### 3.2主要问题

由上可见，KNN算法主要包含三个问题：决定距离度量的方式；决定聚类数目k；决定分类决策规则。

#### 3.2.1距离度量

距离度量反映了两个数据之间的相似程度。

距离：，其中l为空间维数

当p为1时，称为曼哈顿距离

当p为2时，称为欧氏距离

#### 3.2.2聚类数目k

当选取较小的k值时，近似误差较小，但估计误差较大，整体模型复杂，容易过拟合；

当选取较大的k值时，估计误差较小，但近似误差较大，整体模型简单，甚至当k为无穷时，相当于数据集的所有数据参与决策，模型过于简单。

一般选择一个较小的k值，采取交叉验证的方法迭代选取最优的k值。

#### 3.2.3分类决策规则

一般采取多数表决规则，即k个近邻数据中类别最多的为数据所属的类别。误分类率为：



这等价于经验风险最小化。

### 3.3 KNN算法实现问题

KNN算法实现起来的主要难度在于如何对数据进行快速的k近邻搜索。当数据集较小时，可以采用线性搜索，即计算每个数据的k近邻时，把所有的数据算一遍。当数据集较大时，就得考虑使用特殊的数据结构来减少计算距离的时间。这里选用kd树数据结构来实现，可以省去对大多数数据的搜索。