kNN:算法

算法思路：如果一个样本在特征空间中的k个最相似(即特征空间中最邻近)的样本中的大多数属于某一个类别，则该样本也属于这个类别。

由于KNN方法主要靠周围有限的邻近的样本，而不是靠判别类域的方法来确定所属类别的，因此对于类域的交叉或重叠较多的待分样本集来说，KNN方法较其他方法更为适合。

不足：1.当样本不平衡时，如一个类的样本容量很大，而其他类样本容量很小时，有可能导致当输入一个新样本时，该样本的K个邻居中大容量类的样本占多数。-解决办法：加权。

2计算量大，

K-NN可以看成：有那么一堆你已经知道分类的数据，然后当一个新数据进入的时候，就开始跟训练数据里的每个点求距离，然后挑离这个训练数据最近的K个点看看这几个点属于什么类型，然后用少数服从多数的原则，给新数据归类。

**算法步骤：**

step.1---初始化距离为最大值

step.2---计算未知样本和每个训练样本的距离dist

step.3---得到目前K个最临近样本中的最大距离maxdist

step.4---如果dist小于maxdist，则将该训练样本作为K-最近邻样本

step.5---重复步骤2、3、4，直到未知样本和所有训练样本的距离都算完

step.6---统计K-最近邻样本中每个类标号出现的次数

step.7---选择出现频率最大的类标号作为未知样本的类标号

k值通常是人为预先定义的常值，从上图可以看出，k值的选取对会对结果有很多的影响。大的k值可以减小噪声对分类的影响，但是和实例较远的实类会对判决起作用，通常采用**交叉验证**的方式来选取最优的k值。（k为不大于20的整数）

最优的K邻域