前面几篇笔记介绍的判别方法主要有：线性回归，回归和，实际上这三种判别都是参数化方法：即判别方法都给出了显示的判别式或者概率分布的具体函数形式，并且这些函数由少量的参数控制，而参数的值可以由数据集确定。而这种方法的一个重要局限性就是选择的概率密度模型可能对于生成数据来说是一个很差的模型，导致预测表现很差。于是，一些非参数化的方法进行概率密度估计就被提出来。主要有：核密度估计和近邻方法。

### 数学思想

假设观测服从维空间的某个未知的概率分布，我们的目的就是要估计的值，考虑一个包含的某个小区域，这个区域的概率质量为：



假设我们收集了服从分布的次观测，每个数据点落在区域中的概率为，因此位于区域内部的数据点的总数将服从二项分布：



可知落在区域内部的数据点的平均比例为：



概率分布方差为：



1）当很大时，分布会在均值附近产生尖峰，并且；

2）同时假定区域足够小，使得在区域内的概率密度大致为常数，则有：

，其中是区域的体积。

则可以得到：



上式的成立依赖于两个相互矛盾的假设：1）区域要足够小，使得这个区域内的概率密度近似常数；2）区域要足够大，使得落在这个区域内的数据点的数量能够足够让二项分布达到尖峰。

如果我们固定的值，利用数据确定，这就是核方法；固定的值，利用数据确定值，这就是近邻方法。

### **KNN算法**

假设有一个数据集，其中个数据点属于类别，数据点的总数为，则。如果我们想对一个新的数据点进行分类，那么我们可以画一个以为中心的球体，这个球体精确的包含个数据点（无论属于哪个类别），假设球体的体积为，并且包含来自类别的个数据点，则：



即如果我们想最小化错误分类的概率，那么我们可以把测试点分配给有最大后验概率的类别，这对应与最大的。因此，为了分类一个新数据点，我们从训练数据中选择个最近的数据点，然后把新的数据点分配为这个集合中数量最多的点的类别。

算法步骤：

1. 计算已知类别数据集中每个点与当前点的距离；
2. 选取与当前点距离最小的个点；
3. 统计前个点中每个类别的样本出现频率；
4. 返回前个点出现频率最高的类别作为当前点的预测分类。

两个问题：

1. 按上述方法，近邻方法需要存储整个训练数据，如果数据集很大的话，存储和计算的代价会很大，可以通过建立一个基于树的搜索结构，使得近邻可以高效的被找到，而不必遍历整个数据集，这个计算可以被抵消，代价就是需要进行一次性的额外计算量，一般使用；
2. 值的选取。选多大的值是个问题，当值取到全部训练数据集时，相当于模型是最简单的，没有考虑到数据集中数据的差异性；当值取很小时，相当于模型复杂，容易出现，值选取一般用交叉验证的方式获得。



2017.08.01.night