**1.K近邻**

**1.1定义**

给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最近邻的k个实例，这k个实例的多数属于某个类，就把该输入实例分为这个类。

**K邻近算法：**

输入：训练数据集

T = {(x1, y1), (x2, y2), ......, (xN, yN)}；

实例特征向量x；

输出：实例x所属的类y。

（1）根据给定的距离度量，在训练集T中找出与x最邻近的k个点，涵盖这k个点的x的邻域记作Nk(x)；

（2）在Nk(x)中根据分类的决策规则（往往是多数表决）决定x的类别y：

其中I为指示函数，i = 1，2，……，N，j = 1，2，……，K。

由算法流程可知，k近邻实际上只是利用训练数据集对特征向量空间进行划分，并作为其分类的“模型”，因此k近邻不具有显式的学习过程，可以认为k邻近是预测时先学习再预测，或者是对训练数据集的预处理即学习的过程。

k近邻优点：精度高、对异常值不敏感、无数据输入假定。

k近邻缺点：计算复杂度高、空间复杂度高、无法给出任何数据的基础结构信息。

**1.2 k近邻的三个基本要素**

**1）距离的度量：**欧式距离及更一般的Lp距离等，不同的距离度量会导致确定的最近邻点是不同的。如果特征的量纲会对距离的度量产生影响，需要在计算距离之前进行归一化或标准化，以消除量纲对距离度量的影响。

**2）k值的选择：**对结果有重大影响。

k值较小时，预测结果会对近邻的实例点非常敏感，优点是“学习”的近似误差（approximation error）会减小，缺点是“学习”的估计误差（estimation error）会增大；k值较大时，与输入实例较远的训练实例也会影响预测，使预测发生错误，优点是估计误差会变小，缺点是近似误差会变大。

k值减小意味着整体模型变得复杂，容易发生过拟合，k值的增大意味着整体模型变得简单。

实际应用时，一般先选取一个较小的k值，然后采用**交叉验证法**来选取最优的k值。

**3）分类决策规则：**往往是多数表决（投票法），由输入实例的k个邻近的训练实例中的多数类决定输入实例的类。多数表决规则等价于经验风险最小化：

假设分类函数为

如果损失函数是0-1损失，误分类的概率是

对给定的输入实例x，其k近邻训练实例集合Nk(x)，涵盖Nk(x)区域的类别是cj，那么误分类率是

由上式可知，要使误分类最小即经验风险最小，就要使最大。

当三个基本要素和训练集确定后，对于任何一个新的输入实例所属的类唯一确定。

**1.3 k近邻搜索**

k邻近在训练实例划分的特征空间中搜索k个邻近点时最简单的方法是线性扫描，但是k近邻的缺点是计算复杂度高，空间复杂度高，尤其当特征空间维度数和训练实例数据量大时，计算非常耗时，所以需要使用特殊的结构存储训练数据，减少计算距离的次数。kd树是其中一个方法。

**1.3.1构造kd树**

输入：k维空间数据集T={x1, x2, ......, xN}，

其中xi = (xi(1), xi(2), ......, xi(k))T，i=1，2，……，N；

输出：kd树。

（1）构造根结点，对应包含T的k维空间的超矩形区域。

（2）递归切分。对深度为j的结点，选择x(l)为切分的坐标轴（l = j(mod k)+1），以该区域中所有实例的x(l)轴的中位数作为切分点，用通过切分点且与x(l)轴垂直的超平面将该区域切分成深度为j+1的左、右子结点，坐标x(l)小于切分点的子区域为左子结点，大于切分点的子区域为右结点，落在切分超平面上的实例点保存在该结点。

（3）终止。两个子区域没有实例时停止。

kd树是二叉树，这样得到的kd树是平衡的，但是平衡kd树搜索时的效率未必是最优的。

**1.3.2 kd树的最近邻搜索**

输入：已构造的kd树；目标点x；

输出：x的最近邻。

（1）在kd树中找出包含目标点x的叶结点：

从根结点出发递归地向下访问kd树，若目标点x当前维的坐标小于切分点的坐标，则移动到左子结点，否则移动到右子结点，直到子节点是叶结点为止。

（2）以此结点为“当前最近点”。

（3）递归地向上回退，在每个结点进行以下操作：

a）如果该结点保存的实例点比当前最近点距离目标点更近，则以该实例点为“当前最近点”；

b）当前最近点一定存在于该结点一个子结点对应的区域，检查该子结点的父结点的另一个字结点对应的区域是否有更近的点。具体地，检查另一子结点对应的区域是否与以目标点为球心、以目标点与“当前最近点”间的距离为半径的超球体相交。

如果相交，可能在另一个子结点对应的区域内存在距目标点更近的点，移动到另一个子结点，接着递归地进行最近邻搜索；如果不相交，向上回退。

（4）当回退到根结点时，搜索结束。最后的“当前最近点”即为x的最邻近点。

kd树更适用于训练实例数远大于空间维度数时的k近邻搜索，当空间维度数接近训练实例数时，它的效率会迅速下降，几乎接近线性扫描。

如果实例点是随机分布的，kd树搜索的平均时间复杂度为O(logN)（N是训练实例数）。