k-nearest Neighbour (kNN，k近邻法)

1. k近邻算法

k近邻算法（近朱者赤，近墨者黑）：给定训练数据集，对新输入的实例，在训练数据集中寻找与该实例最近邻的k个实例，这k个实例多属于某个类，则将输入实例分到这个类别中。

算法1（k近邻法）

输入：训练数据集

$$T=\left\{\left(x_1, y_1\right),\left(x_2, y_2\right), \cdots,\left(x_N, y_N\right)\right\}$$

其中，$x_i \in \mathcal{X} \subseteq \mathbf{R}^n$为实例的特征向量，$y_i \in \mathcal{Y}=\{c_1, c_2, \cdots, c_K\}$为实例的类别，$i=1,2,...,N$；实例特征向量$x$；

输出：实例x所属的类别y

（1）根据给定的距离度量，在训练集T中找出与x最近邻的k个点，涵盖这k个点的x的邻域记作$N_k(x)$；

（2）在$N_k(x)$中根据分类决策规则（如多数表决）决定x的类别y：

$$y=\arg \max _{c_j} \sum_{x_i \in N_k(x)} I\left(y_i=c_j\right), \quad i=1,2, \cdots, N ; \quad j=1,2, \cdots, K \tag{1.1}$$

式（1.1）中，I为示性函数，即当$y_i=c_j$时I为1，否则为0。

2. k近邻模型

k近邻法中，当(a) 训练集 (b) 距离度量 (c) k值 (d) 决策规则确定后，分类唯一确定。 相当于将特征空间划分成一些子空间，确定子空间每个点所属的类别。

2.1. 距离度量

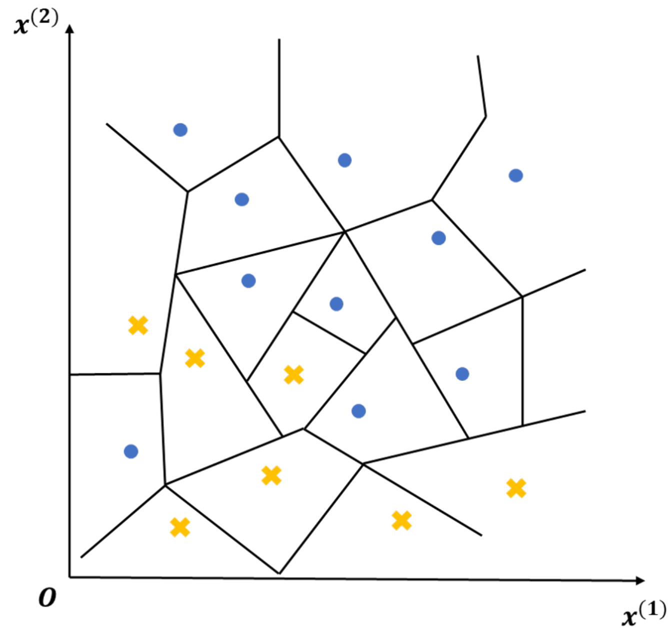

Figure 1: k近邻法的模型对应特征空间的一个划分

特征空间中实例点之间的距离是其相似度的体现，k近邻模型的特征空间一般是n维实数向量空间$\mathbb{R}^n$，除了欧氏距离外，一般也是用L_p距离或Minkowski距离。

设特征空间$\mathcal{X}$是n维实数向量空间$\mathbb{R}^n, x_i, \quad x_j \in \mathcal{X}, x_i = \left(x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}\right)^\top, x_j = \left(x_j^{(1)}, x_j^{(2)}, \cdots, x_j^{(n)}\right)^T$，$x_i, x_j$的$L_p$距离定义为

$$L_p\left(x_i, x_j\right) = \left(\sum_{l=1}^{n}\left|x_i^{(l)} - x_j^{(l)}\right|^p\right)^{\frac{1}{p}} \tag{2.1}$$

这里$p \geq 1$。当$p = 2$时，称为欧氏距离（Euclidean distance），即

$$L_2\left(x_i, x_j\right) = \left(\sum_{l=1}^{n}\left|x_i^{(l)} - x_j^{(l)}\right|^2\right)^{\frac{1}{2}} \tag{2.2}$$

当$p = 1$时，称为曼哈顿距离（Manhattan distance），即

$$L_1\left(x_i, x_j\right) = \sum_{l=1}^{n}\left|x_i^{(l)} - x_j^{(l)}\right| \tag{2.3}$$

当$p=\infty$时，它是各个坐标距离的最大值，即

$$L_{\infty}\left(x_i, x_j\right)=\max_{l}\left|x_i^{(l)}-x_j^{(l)}\right| \tag{2.4}$$

2.2. k值的选择

k值的选择会对最后分类结果产生重大影响。

若k值较小，则此时使用较小的邻域进行预测，结果就会对最近邻点比较敏感，易发生过拟合。

若k值较大，与输入实例较远的x也会预测起到作用，可能造成较大的近似误差。特例：当k = N时，相当于求类别均值。

一般通过交叉验证法选取k值。

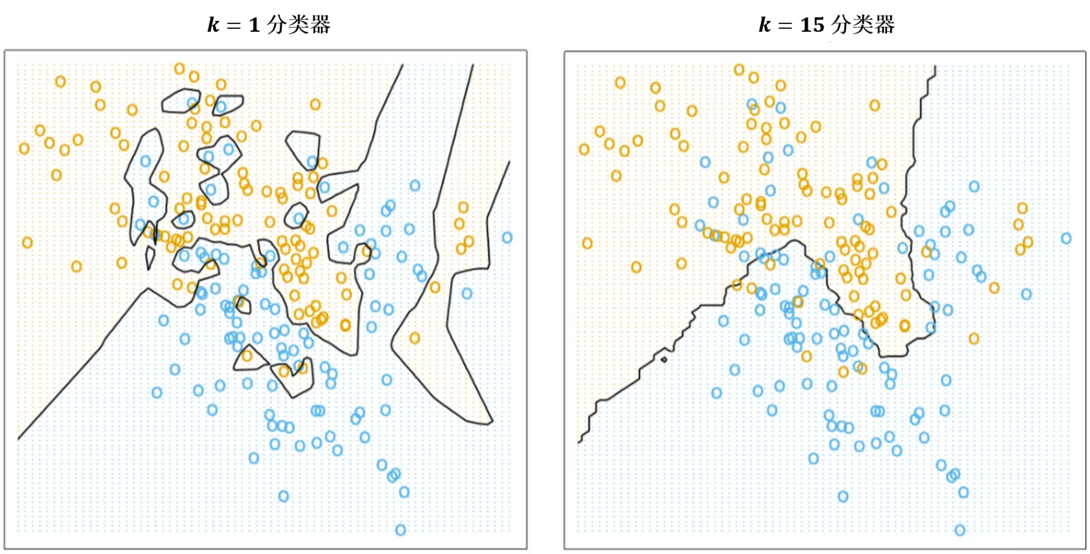

Figure 2: 不同k值对应的k近邻分类器

2.3. 分类决策规则

一般选择“投票法”（majority voting rule）。

解读：多数表决法等价于经验损失最小化（0-1损失）

设分类函数为 $f : \mathbb{R}^p \to \{c_1, \cdots, c_K\}$. 则误分类概率为：

$$P(Y \neq f(X)) = 1 - P(Y = f(X)). \tag{2.5}$$

对于给定的实例 x, 最近的k个邻居构成集合$N_k(x)$。假设涵盖$N_k(x)$ 的类别是c_j，则该邻域内误分类率（经验风险）为

$$\frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i \neq c_j) = 1 - \frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i = c_j)$$

要使经验风险最小，则需要使$\sum_{x_i \in N_k(x)} I(y_i = c_j)$最大。因此，应该选取$c_j$为票数最多的类。