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。

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。

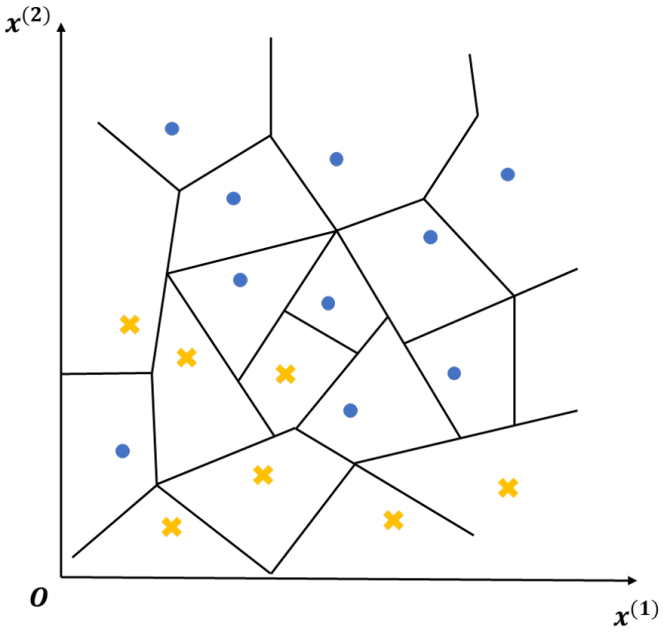

图 1: k 近邻法的模型对应特征空间的一个划分

2.1. 距离度量

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。

2.3. 分类决策规则

一般选择“投票法”（majority voting rule）。

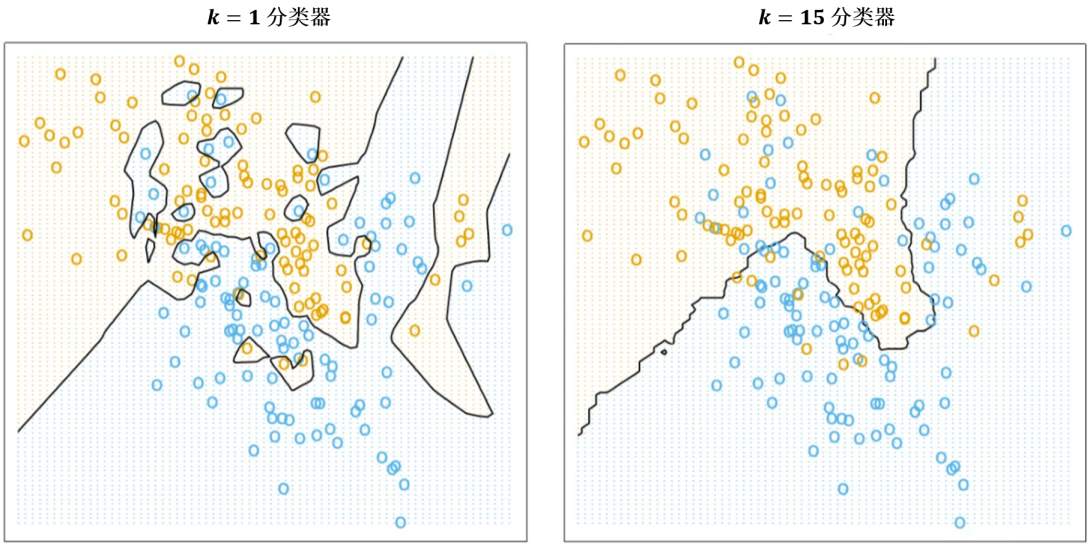

图 2: 不同 k 值对应的 k 近邻分类器