KNN（k邻近法）

1. K邻近算法

输入：训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$

其中，$x_i\in\mathcal{X}\in R^n$ 是实例的特征向量，$y_i\in\mathcal{Y}=\{c_1,c_2,\cdots,c_K\}$ 为实例的类别，

$i=1,2,\cdots,N$；实例特征向量 x_i

(1) 根据给定的度量距离，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 $N_k(x)$

(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

$$y=\arg\max_{c_j}\sum_{x_j\in N_k(x)} I(y_i=c_j),\ i=1,2,\cdots,N;\ j=1,2,\cdots,K$$

2. K邻近模型

2.1 距离度量

设特征空间 $\mathcal{X}$ 是n维实数向量空间 R^n，$x_i,x_j\in\mathcal{X}$，x_i,x_j 的 L_p 距离定义为

$$L_p(x_i,x_j)=\left(\sum_{l=1}^{n}|x_i^{(l)}-x_j^{(l)}|^p\right)^{1/p},\ \text{其中 } p\geq 1$$

$p=1$ 时称为曼哈顿距离（Manhattan Distance）

$p=2$ 时称为欧氏距离（Euclidean Distance）

$p=\infty$ 时，$L_\infty(x_i,x_j)=\max\limits_{l}|x_i^{(l)}-x_j^{(l)}|$

2.2. K值的选择

若k值较小，易发生过拟合；若k值较大，可能造成较大近似误差

一般使用交叉验证法选取k的值

2.3 分类决策规则

一般选取"投票法"（majority voting rule）

解读：多数投票法等价于经验损失最小化（0-1损失）

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

最大化 $\sum_{x_i\in N_k(x)} I(y_i=c_j)$ 等价于最小化 $\frac{1}{k}\sum_{x_i\in N_k(x)} I(y_i\neq c_j)$