

K-Nearest Neighbor 推导

Soul Walker

2020 年 7 月 22 日

输入：训练数据集

$$T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_N, y_N)\} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$$

其中， $\mathbf{x}_i \in \chi \in \mathbb{R}^N$ 为实例的特征向量， $\mathbf{x}_i = (\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(n)})^T$ ， $y_i \in \{+1, -1\}$ 为实例的类别， $i = 1, 2, \dots, N$ ，实例特征向量 \mathbf{x} ， \mathbf{X} 为训练样本集，形状为 (N, p) ， \mathbf{Y} 是训练标签集，形状为 $(N, 1)$ 。

输出：实例 \mathbf{x} 所属的类别 y 。

我们首先给出 x_i, x_j 的 L_p 距离定义：

$$L_p(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^n |\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}|^p \right)^{\frac{1}{p}}$$

此处所使用的是**欧氏距离**即当 $p = 2$ 时，

$$L_2(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^n |\mathbf{x}_i^{(l)} - \mathbf{x}_j^{(l)}|^2 \right)^{\frac{1}{2}}$$

KNN 算法所使用的决策规则为**多数表决规则 (Majority Voting Rule)**，即在输入的实例的 k 个临近的训练实例中由出现次数最多的类表示表示最终结果。