陕西科技大学
中国 西安 710021
SHAANXI UNIVERSITY OF SCIENCE & TECHNOLOGY
XI' AN 710021, P. R. CHINA

机器学习 - KNN学习笔记

一、模型

三要素：① 距离度量 ② K-值的选择 ③ 分类决策规则

利用训练数据集对特征向量空间进行划分

二. 距离度量

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

② 欧氏距离：$L_2(x_i, x_j) = \left(\sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|^2\right)^{\frac{1}{2}}$

③ 曼哈顿距离 $L_1(x_i, x_j) = \sum_{l=1}^{n} |x_i^{(l)} - x_j^{(l)}|$

④ 切比雪夫距离 $L_\infty(x_i, x_j) = \max_l |x_i^{(l)} - x_j^{(l)}|$

eg. $x_1 = (1,1)^T$ $x_2 = (5,1)^T$ $x_3 = (4,4)^T$

当 $p=1$ 时：$L_1(x_1, x_2) = 4$ $L_2(x_1, x_3) = 6$ 故 x_2 为最近邻点

当 $p=2$ 时 $L_2(x_1, x_2) = 4$ $L_2(x_1, x_3) = 4.24$ 故 x_2 为最近邻点

当 $p=3$ 时 $L_3(x_1, x_2) = 4$ $L_3(x_1, x_3) = 3.78$ 故 x_3 为最近邻点

当 $p=4$ 时 $L_4(x_1, x_2) = 4$ $L_4(x_1, x_3) = 3.57$ 故 x_3 为最近邻点

当 $p>4$ 时 $L_p(x_1, x_2) = 4$ $L_p(x_1, x_3) < 4$ 故 x_3 为最近邻点

当p=∞时 $L_\infty(x_1,x_2)=4$ $L_\infty(x_1,x_2)=3$ 的 x_2 为最近邻点

综上：较小k值：学习近似误差↓ 但估计误差↑ 敏感性↑ 模型复杂

较大k值：估计误差↓ 近似误差↑

k的取值可以交叉验证选取，依于样本量的平方根

三、分类决策规则

多数表决规则：由输入实例的k个邻近的训练实例中的数类决定输入实例的类。

分类函数：$f: R^n \to \{c_1, c_2, \cdots c_k\}$

0-1损失函数：$L(Y, f(x)) = \begin{cases} 1, & Y \neq f(x) \\ 0, & Y = f(x) \end{cases}$

误分类概率：$P(Y \neq f(x)) = 1 - P(Y = f(x))$

$x \in X$, $\frac{1}{k}\sum_{x_i \in N_k(x)} I(y_i \neq c_i) = 1 - \frac{1}{k}\sum_{D_i \in N_k(x)} I(y_i = c_i)$

$\arg\max \sum_{x_i \in N_k(x)} I(y_i = c_j)$

KNN 应对分类回归问题