

机器学习基础作业 2

原梓轩

2025 年 3 月 26 日

问题 1. 把对缺失值的处理推广到 *Gini* 指数的计算.

证明. 给定训练集 D 和属性 a , 令 \tilde{D} 表示 D 中在属性 a 上没有缺失值的样本子集. 假定属性 a 有 V 个可取值 $\{a^1, a^2, \dots, a^V\}$, 令 \tilde{D}^v 表示 \tilde{D} 中在属性 a 上取值为 a^v 的样本子集, \tilde{D}_k 表示 \tilde{D} 中属于第 k 类 ($k = 1, 2, \dots, |\mathcal{Y}|$) 的样本子集.

为每个样本 \mathbf{x} 赋予一个权重 $w_{\mathbf{x}}$, 并定义

$$\rho = \frac{\sum_{\mathbf{x} \in \tilde{D}} w_{\mathbf{x}}}{\sum_{\mathbf{x} \in D} w_{\mathbf{x}}}$$
$$\tilde{r}_v = \frac{\sum_{\mathbf{x} \in \tilde{D}^v} w_{\mathbf{x}}}{\sum_{\mathbf{x} \in \tilde{D}} w_{\mathbf{x}}} \quad (1 \leq v \leq V)$$
$$\tilde{p}_{vk} = \frac{\sum_{\mathbf{x} \in \tilde{D}_k^v} w_{\mathbf{x}}}{\sum_{\mathbf{x} \in \tilde{D}^v} w_{\mathbf{x}}} \quad (1 \leq k \leq |\mathcal{Y}|)$$

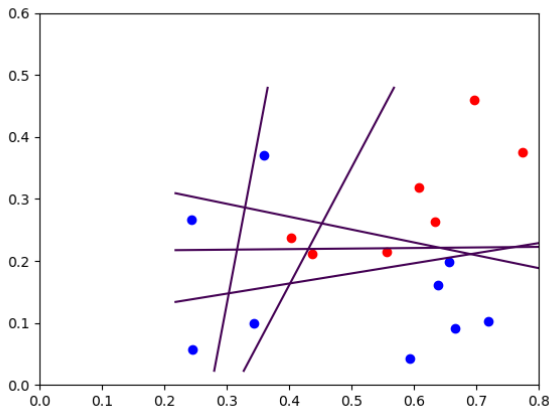
基于以上定义, 可以推广 *Gini* 指数的计算为

$$\text{Gini}(D, a) = \rho \times \text{Gini}(\tilde{D}, a) = \rho \times \sum_{v=1}^V \tilde{r}_v \times \text{Gini}(\tilde{D}^v) = \rho \times \sum_{v=1}^V \tilde{r}_v \times \left(1 - \sum_{k=1}^{|\mathcal{Y}|} \tilde{p}_{vk}^2\right)$$

□

问题 2. 下载或编程实现多变量决策树算法, 并在西瓜数据集 3.0 上进行测试.

证明. 编程实现在随同的 `decision_tree.py` 已经实现, 测试结果如下:



红蓝颜色的点表示好坏西瓜, 直线表示决策树的划分边界.

□