assignment3

Xiaoma

2022 年 10 月 23 日

题目 1.

解: 决策树递归返回条件为：

- 当前结点包含的样本全部属于同一类别
- 当前属性集为空或所有样本在所有属性集取值相同
- 当前结点包含的样本集合为空

已知不含冲突数据，即不包含不能划分的数据，必存在训练误差为 0 的决策树。

题目 2.

解: 数据集 D 的基尼值

$$Gini(D) = \sum_{k=1}^{|\mathrm{y}|} p_k(1 - p_k)$$

属性 α 的基尼指数

$$Gini - index(D, \alpha) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v)$$

$\tilde{D}$ 表示在属性 α 上没有缺失值的样本子集，$\tilde{D}^v$ 表示 $\tilde{D}$ 中在属性 α 上取值为 α^v 的样本子集，$\tilde{D}_k$ 表示 $\tilde{D}$ 中属于第 k 类的样本子集，w_x 为每个样本的权重。

$$\mathrm{Ent}(\widetilde{D}) = -\sum_i \tilde{p}_k \log_2 \tilde{p}_k$$

$$\tilde{p}_k = \frac{\sum_{x\in\widetilde{D}_k} w_x}{\sum_{x\in\widetilde{D}} w_x}$$

$$\tilde{r}_v = \frac{\sum_{x\in\widetilde{D}^v} w_x}{\sum_{x\in\widetilde{D}} w_x}$$

$$\rho = \frac{\sum_{x\in\widetilde{D}} w_x}{\sum_{x\in D} w_x}$$

$$\mathrm{Gini-index}(D,a) = \rho\sum_{v=1}^{V}\tilde{r}_v\,\mathrm{Gini}\left(\widetilde{D}^v\right)$$

题目 3.

解:

$$L(\boldsymbol{p},\lambda) = -\sum_{i=1}^{K} p_i \log_2 p_i + \lambda(\sum_{i=1}^{K} p_i - 1)$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{K} p_i - 1 = 0$$

$$\frac{\partial L}{\partial p_i} = -\log_2 p_i - \frac{1}{\ln 2} + \lambda = 0$$

$$p_1 = p_2 = ... = p_k = \frac{1}{K}$$

该点即为 $L(\boldsymbol{p},\lambda)$ 在附加条件下的可能极值点，该点只有一个，则可证熵的最大分布为均匀分布。

题目 4.

解:

a.

$$Ent(D) = -\sum_k p_k \log_2 p_k = 1$$

b. 对于属性 A

$$Ent(D^T) = -\sum_k p_k^T \log_2 p_k^T = 0.811$$

$$Ent(D^F) = -\sum_k p_k^F \log_2 p_k^F = 0.918$$

$$Gain(D, A) = Ent(D) - \sum_v \frac{|D^v|}{|D|} Ent(D^v) = 0.125$$

对于属性 B

$$Ent(D^T) = -\sum_k p_k^T \log_2 p_k^T = 0.971$$

$$Ent(D^F) = -\sum_k p_k^F \log_2 p_k^F = 0.971$$

$$Gain(D, B) = Ent(D) - \sum_v \frac{|D^v|}{|D|} Ent(D^v) = 0.029$$

c.

$$T_C = \frac{a^1 + a^2}{2} = 1.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0.108$$

$$T_C = \frac{a^2 + a^3}{2} = 2.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0.236$$

$$T_C = \frac{a^3 + a^4}{2} = 3.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0.035$$

$$T_C = \frac{a^4 + a^5}{2} = 4.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0.125$$

$$T_C = \frac{a^5 + a^6}{2} = 5.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0$$

$$T_C = \frac{a^6 + a^7}{2} = 6.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0.035$$

$$T_C = \frac{a^7 + a^8}{2} = 7.5, Gain(D, C) = Ent(D) - \sum_v \frac{|D_t^\lambda|}{|D|} Ent(D_t^\lambda) = 0.108$$

d.

$$Gini(D) = \sum_{k=1}^{|y|} p_k(1-p_k)$$

$$Gini - index(D,A) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) = 0.417$$

$$Gini - index(D,B) = \sum_{v=1}^{V} \frac{|D^v|}{|D|} Gini(D^v) = 0.48$$

$Gini - index(D,A) < Gini - index(D,B)$，A 比 B 更可取。

e.

$$Gain_rate(D,A) = 0.128$$

$$Gain_rate(D,B) = 0.128$$