Tania Soutonglang

CS 422-01

Sept. 20, 2022

Homework 2

Chapter 3.7

6.

a. Calculate the Gini index and misclassification error rate of the parent node P.

$$\text{Gini}(\text{P}) = 1 - \left(\left(\frac{7}{10}\right)^2 + \left(\frac{3}{10}\right)^2\right)$$
$$= 1 - (0.49 + 0.09)$$
$$= 1 - 0.58$$
$$= \mathbf{0.42}$$

$$\text{Error}(\text{P}) = 1 - \max\left(\frac{7}{10}, \frac{3}{10}\right)$$
$$= 1 - \max(0.7, 0.3)$$
$$= 1 - 0.7$$
$$= \mathbf{0.3}$$

b. Calculate the weighted Gini index of the child nodes. Would you consider this attribute test condition if Gini is used as the impurity measure?

$$\text{Gini}(\text{C1}) = 1 - \left(\left(\frac{3}{3}\right)^2 + \left(\frac{0}{3}\right)^2\right)$$
$$= 1 - (1 + 0)$$
$$= 1 - 1$$
$$= \mathbf{0}$$

$$\text{Gini}(\text{C2}) = 1 - \left(\left(\frac{4}{7}\right)^2 + \left(\frac{3}{7}\right)^2\right)$$
$$= 1 - (0.33 + 0.18)$$
$$= 1 - 0.51$$
$$= \mathbf{0.49}$$

I would use this attribute since C1 has a Gini index of 0, indicating perfect equality, and C2 has a Gini index of 0.49 which is in the middle of perfect equality and perfect inequality.

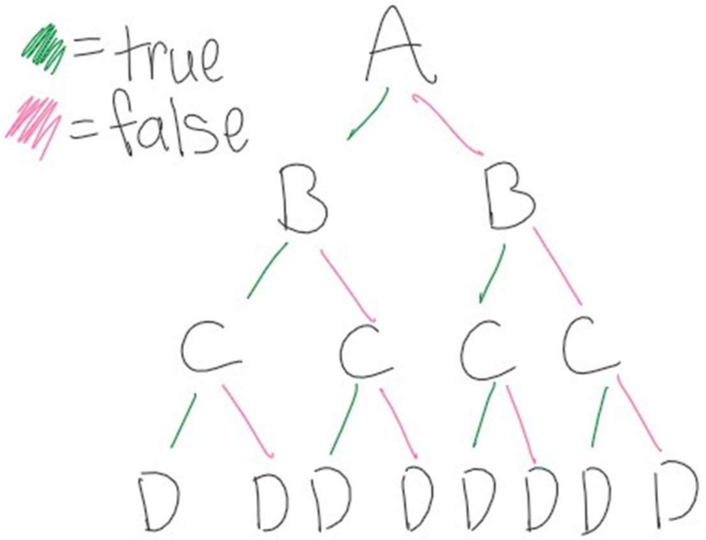
c. Calculate the weighted misclassification rate of the child nodes. Would you consider this attribute test condition if misclassification rate is used as the impurity measure?

$$\text{Error}(\text{C}1) = 1 - \max\left(\frac{3}{3}, \frac{0}{3}\right)$$
$$= 1 - \max(1{,}0)$$
$$= 1 - 1$$
$$= \mathbf{0}$$

$$\text{Error}(C2) = 1 - \max\left(\frac{4}{7}, \frac{3}{7}\right)$$
$$= 1 - \max(0.57, 0.43)$$
$$= 1 - 0.57$$
$$= \mathbf{0.43}$$

I would consider this attribute test condition since the misclassification rate for C1 is 0 so it is perfectly classified.

1. Draw the full decision tree for the parity function of four Boolean attributes, A, B, C, and D. Is it possible to simplify the tree?
   It is not possible to simplify the tree because the only outputs each attribute can have is “true” or “false”, so there is no other way to simplify the outputs even further.



3.

a. What is the entropy of this collection of training examples with respect to the class attribute?

Target Class

+	-
4	5

$$E(Total) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = \mathbf{0.9911}$$

b. What are the information gains of a1 and a2 relative to these training examples?

Column a1

	+	-
T	3	1
F	1	4

$$E(a1) = \frac{4}{9}(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}) + \frac{5}{9}(-\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5})) = 0.7616$$

$$\text{information gain (a1)} = 0.9911 - 0.7616 \approx \mathbf{0.23}$$

Column a2

	+	-
T	2	3
F	2	2

$$E(a2) = \frac{4}{9}(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}) + \frac{5}{9}(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4})) = 0.9871$$

$$\text{information gain (a2)} = 0.9911 - 0.9871 \approx \mathbf{0.004}$$

c. For a3, which is a continuous attribute, compute the information gain for every possible split.

instance	a3	class	split-point	entropy	information gain
1	1.0	+			
6	3.0	-	2.0	0.8484	0.1427
4	4.0	+	3.5	0.9885	0.0026
3	5.0	-	4.5	0.9183	0.0728
9	5.0	-	0.0	0.9839	0.0072
2	6.0	+	5.5	0.9839	0.0072
5	7.0	-	6.5	0.9728	0.0183
8	7.0	+	0.0	0.9728	0.0183
7	8.0	-	7.5	0.8889	0.1022

d. What is the best split (among a1, a2 and a3) according to the information gain?

a1 would have the best split among a1, a2, and a3 since it has the highest information gain.

e. What is the best split (between a1 and a2) according to the misclassification error rate?

f. What is the best split (between a1 and a2) according to the Gini index?

$$\text{Gini(a1)} = \frac{4}{9}\left(1 - \frac{3^2}{4} - \frac{1^2}{4}\right) + \frac{5}{9}\left(1 - \frac{1^2}{5} - \frac{4^2}{5}\right)$$
$$= \mathbf{0.3444}$$

$$\text{Gini(a2)} = \frac{4}{9}\left(1 - \frac{2^2}{4} - \frac{2^2}{4}\right) + \frac{5}{9}\left(1 - \frac{3^2}{5} - \frac{2^2}{5}\right)$$
$$= \mathbf{0.4889}$$

The Gini index of a1 is smaller than a2 so a1 has the best split.

4. Show that the entropy of a node never increases after splitting it into smaller successor nodes.

$$E = \sum_{\text{j=1}}^{\text{n}} p_j \log_2 p_j$$
$$n = C = \{c_1, c_2, \dots, c_n\}$$
$$m = D = \{d_1, d_2, \dots, d_n\}$$

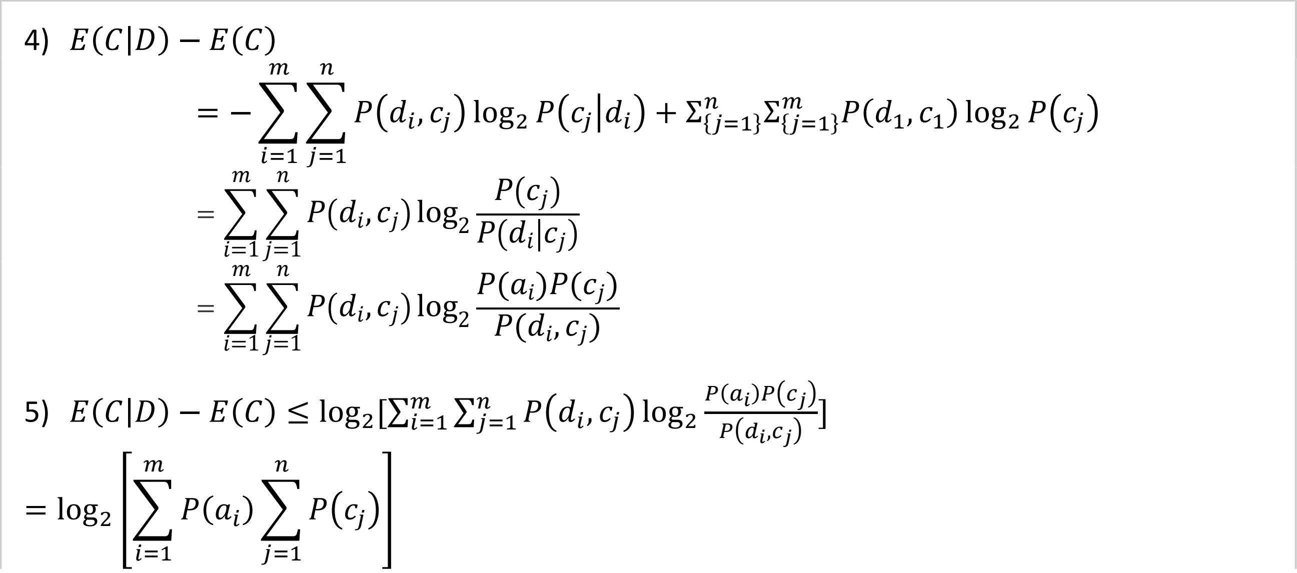
1) $E(C) = \Sigma_{\{j=1\}}^{n} P(c_1) \log_2 P(c_j)$

$= \Sigma_{\{j=1\}}^{n} \Sigma_{\{j=1\}}^{m} P(d_1, c_1) \log_2 P(c_j)$

2) $E(C|d_i) = -\sum_{j=1}^{n} P(c_j|d_i) \log_2 P(c_j|d_i)$

3) $E(C|D) = \sum_{i=1}^{m} P(d_i) E(C|d_i)$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{n} P(a_i) P(c_j|d_i) \log_2 P(c_j|d_i)$$
$$= -\sum_{i=1}^{m}\sum_{j=1}^{n} P(d_i, c_j) \log_2 P(c_j|d_i)$$



4) $E(C|D) - E(C)$

$$= -\sum_{i=1}^{m}\sum_{j=1}^{n} P(d_i, c_j) \log_2 P(c_j|d_i) + \Sigma_{\{j=1\}}^{n} \Sigma_{\{j=1\}}^{m} P(d_1, c_1) \log_2 P(c_j)$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} P(d_i, c_j) \log_2 \frac{P(c_j)}{P(d_i|c_j)}$$
$$= \sum_{i=1}^{m}\sum_{j=1}^{n} P(d_i, c_j) \log_2 \frac{P(a_i)P(c_j)}{P(d_i, c_j)}$$

5) $E(C|D) - E(C) \le \log_2[\sum_{i=1}^{m} \sum_{j=1}^{n} P(d_i, c_j) \log_2 \frac{P(a_i)P(c_j)}{P(d_i,c_j)}]$

$$= \log_2\left[\sum_{i=1}^{m} P(a_i) \sum_{j=1}^{n} P(c_j)\right]$$

$$= \log_2(1)$$
$$= 0$$

Since $E(C|D) - E(C) \leq 0$, entropy does not increase after splitting an attribute.