SC4000/CZ4041/CE4041: Machine Learning

Solutions to L5 Tutorial Questions

Kelly KE

School of Computer Science and Engineering,

NTU, Singapore

Entropy & Information Gain

$$\text{Entropy}(t) = -\sum_{c} P(y = c; t) \log_2 P(y = c; t)$$

- Suppose a parent node t is split into P partitions (children)
- Information Gain:

$$\Delta_{\text{info}} = \text{Entropy(t)} - \sum_{j=1}^{P} \frac{n_j}{n} \text{Entropy}(j)$$

Number of examples at child j

Number of examples at node t

- To choose a feature whose test condition maximizes the gain

Question 1.1

Table 1: Data set for Question 1.

A	B	Class Label
M	F	+
F	T	+
T	T	+
M	F	-
M	F	-
F	F	-
N	F	-
N	T	-
T	T	-
T	F	-

Task: calculate information gain when splitting on A (multi-way) and B. Which feature to choose?

	Parent
+	3
−	7

	$A = T$	$A = F$	$A = M$	$A = N$
+	1	1	1	0
−	2	1	2	2

Split on A

	$B = T$	$B = F$
+	2	1
−	2	5

Split on B

$$\text{Entropy(Parent)} = -\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{7}{10}\right)\log_2\left(\frac{7}{10}\right) = 0.8813$$

	Parent
+	3
−	7

$$\text{Entropy}(A = T) = -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) = 0.9183$$

$$\text{Entropy}(A = F) = -\left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\log_2\left(\frac{1}{2}\right) = 1$$

	$A = T$	$A = F$	$A = M$	$A = N$
+	1	1	1	0
−	2	1	2	2

$$\text{Entropy}(A = M) = -\left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) - \left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) = 0.9183$$

Split on *A*

$$\text{Entropy}(A = N) = -\left(\frac{0}{2}\right)\log_2\left(\frac{0}{2}\right) - \left(\frac{2}{2}\right)\log_2\left(\frac{2}{2}\right) = 0$$

$$\text{Entropy}(\text{Split}_A) = \left(\frac{3}{10}\right) \times 0.9183 + \left(\frac{2}{10}\right) \times 1 + \left(\frac{3}{10}\right) \times 0.9183 + \left(\frac{2}{10}\right) \times 0 = 0.7510$$

$$\Delta_{\text{info}}(A) = 0.8813 - 0.7510 = 0.1303$$

$$\text{Entropy(Parent)} = -\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{7}{10}\right)\log_2\left(\frac{7}{10}\right) = 0.8813$$

	Parent
+	3
−	7

$$\text{Entropy}(B = T) = -\left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\left(\frac{2}{4}\right) = 1$$

$$\text{Entropy}(B = F) = -\left(\frac{1}{6}\right)\log_2\left(\frac{1}{6}\right) - \left(\frac{5}{6}\right)\log_2\left(\frac{5}{6}\right) = 0.65$$

	$B = T$	$B = F$
+	2	1
−	2	5

Split on B

$$\text{Entropy}(\text{Split}_B) = \left(\frac{4}{10}\right) \times 1 + \left(\frac{6}{10}\right) \times 0.65 = 0.79$$

$$\Delta_{\text{info}}(B) = 0.8813 - 0.79 = 0.0913 \quad < \quad \Delta_{\text{info}}(A) = 0.1303$$ ✓

Question 1.2: Calculate Gain Ratio

- Suppose a parent node t is split into P partitions (children)

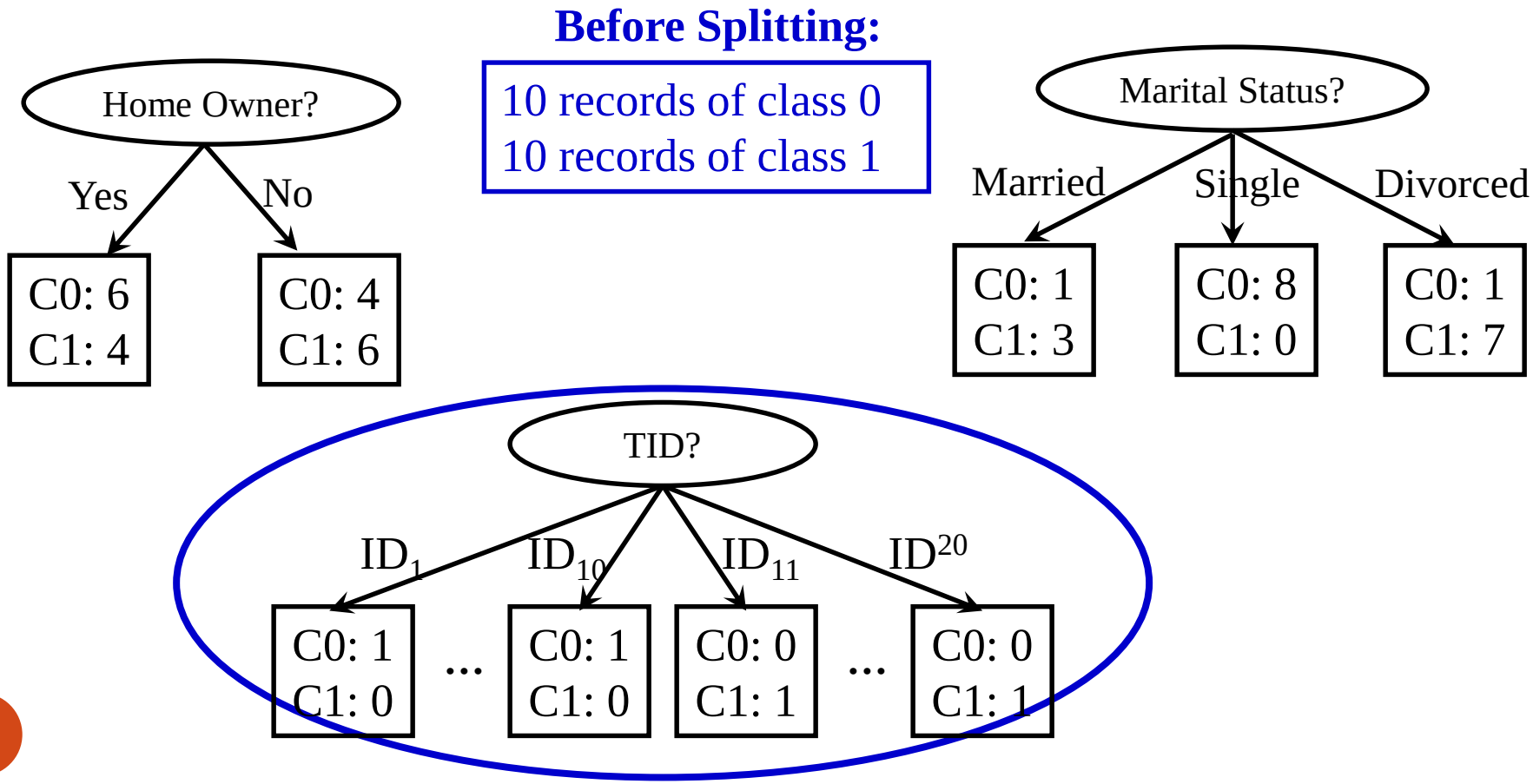
$$\text{Gain Ratio} = \frac{\Delta_{\text{info}}}{\text{SplitINFO}}$$

$$\text{where SplitINFO} = -\sum_{j=1}^{P} \frac{n_j}{n} \log_2 \frac{n_j}{n}$$

The number of records in partition j

Why Gain Ratio?

- Disadvantage: tends to prefer splits that result in large number of partitions, each being small but pure



Question 1.2

$$\Delta_{\text{info}}(A) = 0.1303 \qquad \Delta_{\text{info}}(B) = 0.0913$$

$$\text{Gain Ratio} = \frac{\Delta_{\text{info}}}{\text{SplitINFO}} \quad \text{where SplitINFO} = -\sum_{i=1}^{P} \frac{n_i}{n} \log_2 \frac{n_i}{n}$$

$A = T$	$A = F$	$A = M$	$A = N$
3	2	3	2

$$\text{SplitINFO}(A) = -\left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right) - \left(\frac{3}{10}\right)\log_2\left(\frac{3}{10}\right) - \left(\frac{2}{10}\right)\log_2\left(\frac{2}{10}\right) = 1.9710$$

$$\text{GainRatio}_A = \frac{\Delta_{\text{info}}(A)}{\text{SplitINFO}(A)} = \frac{0.1303}{1.9710} = 0.0661$$

$B = T$	$B = F$
4	6

$$\text{SplitINFO}(B) = -\left(\frac{4}{10}\right)\log_2\left(\frac{4}{10}\right) - \left(\frac{6}{10}\right)\log_2\left(\frac{6}{10}\right) = 0.9710$$

$$\text{GainRatio}_B = \frac{\Delta_{\text{info}}(B)}{\text{SplitINFO}(B)} = \frac{0.0913}{0.9710} = 0.094$$

Thank you!