CSC4008 homework4: Decision Tree

Name: 樊祥　　Soudent ID: 120090414

T1,

Outlook case:

$$H_P = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.9403$$

$$H_{C_1} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$H_{C_2} = 0$$

$$H_{C_3} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$$

$$H_C = \frac{5}{14}\times 0.971 + \frac{5}{14}\times 0.971 = 0.6936$$

∴ information gain $= H_P - H_C = 0.2467$

$$GINI_P = 1-\left(\left(\frac{9}{14}\right)^2+\left(\frac{5}{14}\right)^2\right) = 0.4592$$

$$GINI_{C_1} = 1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right) = 0.48$$

$$GINI_{C_2} = 1-1^2 = 0$$

$$GINI_{C_3} = 1-\left(\left(\frac{2}{5}\right)^2+\left(\frac{3}{5}\right)^2\right) = 0.48$$

$$GINI_C = \frac{5}{14}\times 0.48 + \frac{5}{14}\times 0.48 = 0.3429$$

∴ GINI index $= 0.4592 - 0.3429 = 0.1163$

Humidity case:

$$H_P = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.9403$$

$$H_{C_1} = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.9852$$

$$H_{C_2} = -\frac{1}{7}\log_2\frac{1}{7} - \frac{6}{7}\log_2\frac{6}{7} = 0.5917$$

$$H_C = \frac{1}{2}\times 0.9852 + \frac{1}{2}\times 0.5917 = 0.7885$$

∴ information gain $= H_P - H_C = 0.1518$

$$GINI_P = 1-\left(\left(\frac{9}{14}\right)^2+\left(\frac{5}{14}\right)^2\right) = 0.4592$$

$$GINI_{C_1} = 1-\left(\left(\frac{3}{7}\right)^2+\left(\frac{4}{7}\right)^2\right) = 0.4898$$

$GINI_{C_2} = 1-((\frac{1}{7})^2+(\frac{6}{7})^2) = 0.2449$

$GINI_C = 0.4898\times\frac{1}{2} + 0.2449\times\frac{1}{2} = 0.3674$

$\therefore$ GINI index $= 0.4592 - 0.3674 = 0.0918$

Wind case:

$H_P = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.9403$

$H_{C_1} = -\frac{3}{6}\log\frac{3}{6} - \frac{3}{6}\log\frac{3}{6} = 1$

$H_{C_2} = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.8113$

$H_C = \frac{6}{14}\times 1 + \frac{8}{14}\times 0.8113 = 0.8922$

$\therefore$ Information gain $= 0.9403 - 0.8922 = 0.0481$

$GINI_P = 1-((\frac{9}{14})^2+(\frac{5}{14})^2) = 0.4592$

$GINI_{C_1} = 1-((\frac{3}{6})^2+(\frac{3}{6})^2) = 0.5$

$GINI_{C_2} = 1-((\frac{6}{8})^2+(\frac{2}{8})^2) = 0.375$

$GZNI_C = \frac{6}{14}\times 0.5 + \frac{8}{14}\times 0.375 = 0.4286$

$\therefore$ GZNI index $= 0.4592 - 0.4286 = 0.0306$

T_2,

$H_P = -\frac{3}{10}\log_2\frac{3}{10} - \frac{7}{10}\log_2\frac{7}{10} = 0.8813$

$GZNI_P = 1-((\frac{3}{10})^2+(\frac{7}{10})^2) = 0.42$

① $H_{C_1} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$

$H_{C_2} = -\frac{1}{6}\log_2\frac{1}{6} - \frac{5}{6}\log_2\frac{5}{6} = 0.65$

$H_C = \frac{4}{10}\times 1 + \frac{6}{10}\times 0.65 = 0.79$

$\therefore$ Information gain $= 0.8813 - 0.79 = 0.0913$

$GINI_{C_1} = 1-((\frac{2}{4})^2+(\frac{2}{4})^2) = 0.5$

$GINI_{c_2} = 1-\left(\left(\frac{1}{6}\right)^2+\left(\frac{5}{6}\right)^2\right) = 0.2778$

$GINI_c = \frac{4}{10}\times 0.5 + \frac{6}{10}\times 0.2778 = 0.3667$

$\therefore$ GINI index $= 0.42 - 0.3667 = 0.0533$

② $H_G = 0$

$H_{G_2} = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$

$H_c = \frac{4}{10}\times 0 + \frac{6}{10}\times 1 = 0.6$

$\therefore$ information gain $= 0.8813 - 0.6 = 0.2813$

$GINI_{c_1} = 1-1^2 = 0$

$GINI_{c_2} = 1-\left(\left(\frac{3}{6}\right)^2+\left(\frac{3}{6}\right)^2\right) = 0.5$

$GINI_c = \frac{4}{10}\times 0 + \frac{6}{10}\times 0.5 = 0.3$

$\therefore$ GINI index $= 0.42 - 0.3 = 0.12$

③ $H_{c_1} = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$H_{c_2} = -\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8} = 0.8113$

$H_c = \frac{2}{10}\times 1 + \frac{8}{10}\times 0.8113 = 0.849$

$\therefore$ information gain $= 0.8813 - 0.849 = 0.0323$

$GINI_{c_1} = 1-\left(\left(\frac{1}{2}\right)^2+\left(\frac{1}{2}\right)^2\right) = 0.5$

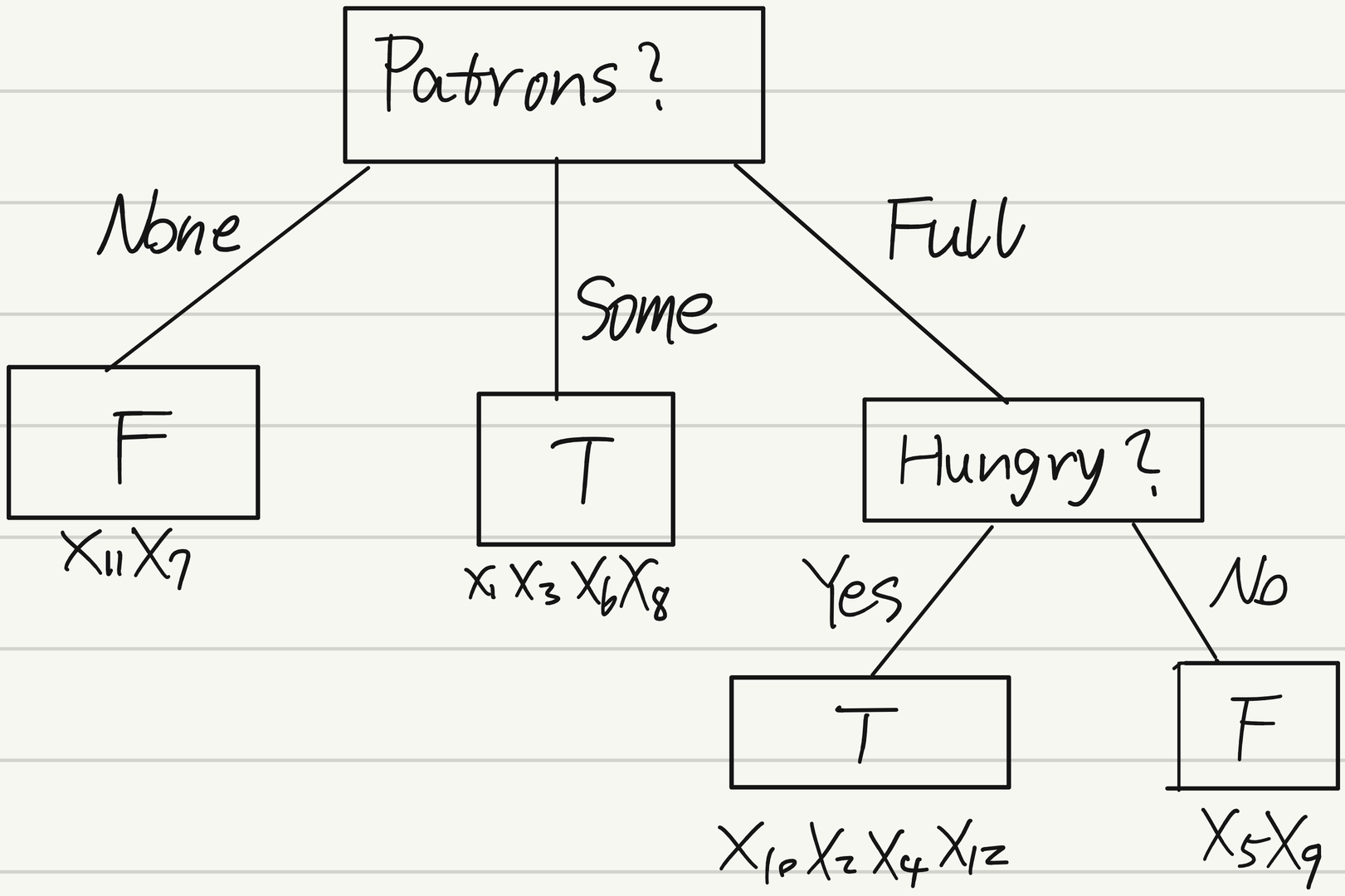
$GINI_{c_2} = 1-\left(\left(\frac{2}{8}\right)^2+\left(\frac{6}{8}\right)^2\right) = 0.375$

$GINI_c = \frac{2}{10}\times 0.5 + \frac{8}{10}\times 0.375 = 0.4$

$\therefore$ GINI index $= 0.42 - 0.4 = 0.02$

since the information gain and gini index in case ② is the largest, so the second way is the best.

T_3.


Patrons?
None
Some
Full
F
$X_{11}X_7$
T
$X_1X_3X_6X_8$
Hungry?
Yes
No
T
$X_{10}X_2X_4X_{12}$
F
X_5X_9