Name = Yanan Liu
NetID = yl2248

P1

Outlook: Gini(S) = $1-(\frac{2}{5})^2-(\frac{3}{5})^2 = 0.48$; Gini(O) = 0; Gini(R) = $1-(\frac{3}{5})^2-(\frac{2}{5})^2 = 0.48$

Gini(outlook) = $\frac{5}{14}\cdot 0.48 + \frac{5}{14}\cdot 0.48 = 0.342$

TEMPER: Gini(H) = $1-(\frac{1}{2})^2-(\frac{1}{2})^2 = \frac{1}{2}$; Gini(M) = $1-(\frac{4}{6})^2-(\frac{2}{6})^2 = \frac{4}{9}$; Gini(C) = $1-(\frac{3}{4})^2-(\frac{1}{4})^2 = 0.375$

Gini(TEMP) = $\frac{4}{14}\cdot\frac{1}{2} + \frac{6}{14}\cdot\frac{4}{9} + \frac{4}{16}\cdot 0.375$

$= 0.427$

Humidity: Gini(H) = $1-(\frac{3}{7})^2-(\frac{4}{7})^2 = 0.489$; Gini(N) = $1-(\frac{1}{7})^2-(\frac{6}{7})^2 = 0.244$

Gini(Humidity) = $\frac{7}{14}\cdot 0.489 + \frac{7}{14}\times 0.244 = 0.366$

Wind: Gini(W) = $1-(\frac{6}{8})^2-(\frac{2}{8})^2 = 0.375$; Gini(S) = $1-(\frac{1}{2})^2-(\frac{1}{2})^2 = \frac{1}{2}$

Gini(wind) = $\frac{8}{14}\cdot 0.375 + \frac{6}{14}\cdot\frac{1}{2} = 0.428$

∴ Gini(outlook) < Gini(Humidity) < Gini(Temp) < Gini(wind)

Root node

Sunny:

TEMPER: Gini(H) = 0; Gini(M) = $\frac{1}{2}$; Gini(C) = 0

Gini(Sunny-Temper) = $\frac{2}{5}\cdot\frac{1}{2} = \frac{1}{5} = 0.2$

Humidity: Gini(H) = 0; Gini(N) = 0

Gini(Sunny-Humidity) = 0

Wind: Gini(W) = $1-\frac{1}{4}-\frac{4}{9} = \frac{4}{9}$; Gini(S) = $1-\frac{1}{4}-\frac{1}{4} = 0.5$

Gini(wind) = $\frac{5}{5}\cdot\frac{4}{9} + \frac{2}{5}\cdot\frac{1}{2} = 0.466$

Rainy:

Temper: Gini(M) = $1-\frac{4}{9}-\frac{1}{9} = 0.444$; Gini(C) = $1-\frac{1}{4}-\frac{1}{4} = 0.5$

Gini(R-T) = $\frac{3}{5}\cdot\ \ +\frac{2}{5}\cdot 0.5 = 0.466$

Humidity: Gini(H) = $1-\frac{1}{4}-\frac{1}{4} = 0.5$; Gini(N) = $1-\frac{4}{9}-\frac{1}{9} = 0.444$

P2:

2.1:

$H(X) = -\sum p_i * \log_2 p_i$

$H(X) = -\frac{4}{15}\log_2\frac{4}{15} - \frac{11}{15}\log_2\frac{11}{15} = 0.97$

2.2:

$H(S2=S) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$

$H(S2=M) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 1$

$H(S2=L) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.917$

$H(S2=XL) = 0$

$H(S2) = \frac{4}{15}\cdot 0.811 + \frac{6}{15}\times 1 + \frac{5}{15}\cdot 0.917 + 0$

$= 0.799$

$IG = 0.97 - 0.799 = 0.170$

$H(S2F=S) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$

$H(S2F=SM) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.917$

$H(\cdots=L) = -\frac{2}{3}\log_2\frac{2}{3} - \frac{1}{3}\log_2\frac{1}{3} = 0.917$

$H(\cdots=XL) = 0$

$H(\cdots=LM) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.917$

$H(S2F) = \frac{4}{15}\cdot 0.811 + \frac{3}{15}\times 0.917 + \frac{3}{15}\times 0.917 + \frac{5}{15}\times 0.917$

$= 0.766$

$IG(S2F) = 0.203$

$IG(S2) < IG(S2F)$

2.3:

Since the equation "$x^2+y^2 \geq 2xy$" $\Rightarrow \frac{x+y}{2} \geq \sqrt{xy}$

the logarithm is monotone increasing

$\therefore \log(\frac{x}{2}+\frac{y}{2}) \geq \log\sqrt{xy} = \log\sqrt{x} + \log\sqrt{y} = \frac{1}{2}\log x + \frac{1}{2}\log y$

$\therefore \log(\frac{x}{2}+\frac{y}{2}) \geq \frac{1}{2}\log x + \frac{1}{2}\log y$.

$\therefore$ when $x=\frac{1}{2}$, the equation holds.

2.4:

let set $f(x) = x\log x$

$f'(x) = \log x + \frac{1}{\ln 2}$

$f''(x) = \frac{1}{x\ln 2}$

$\because x>0 \quad \therefore f''(x) > 0$

$\therefore f(x)$ is convex

$x_1\log\frac{x_1}{y_1} + x_2\log\frac{x_2}{y_2} = y_1\cdot\frac{x_1}{y_1}\log\frac{x_1}{y_1} + y_2\cdot\frac{x_2}{y_2}\log\frac{x_2}{y_2}$

$= y_1 f(\frac{x_1}{y_1}) + y_2 f(\frac{x_2}{y_2})$

$= (y_1+y_2)\left[\frac{y_1}{y_1+y_2} f(\frac{x_1}{y_1}) + \frac{y_2}{y_1+y_2} f(\frac{x_2}{y_2})\right]$

From Jensen's equation:

$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2), \ 0\leq t\leq 1$ [From wikipedia]

$\therefore x_1\log\frac{x_1}{y_1} + x_2\log\frac{x_2}{y_2} = (y_1+y_2)\left[\frac{y_1}{y_1+y_2} f(\frac{x_1}{y_1}) + \frac{y_2}{y_1+y_2} f(\frac{x_2}{y_2})\right]$

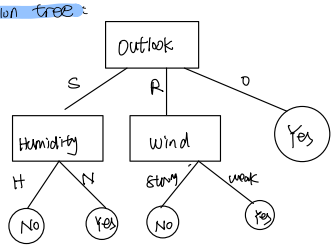
$\geq (y_1+y_2) f\left(\frac{y_1}{y_1+y_2}\cdot\frac{x_1}{y_1} + \frac{y_2}{y_1+y_2}\cdot\frac{x_2}{y_2}\right)$

$= y_1+y_2\, f(\frac{x_1+x_2}{y_1+y_2}) = (x_1+x_2)\log\frac{x_1+x_2}{y_1+y_2}$

$\therefore x_1\log\frac{x_1}{y_1} + x_2\log\frac{x_2}{y_2} \geq (x_1+x_2)\log\frac{x_1+x_2}{y_1+y_2}$ proved.

Gini(R-H) = $1 - \frac{2}{3} + 0 = \frac{3}{5} = 0.400$

Decision tree:



P2:

3.1

Entropy:

$H(X) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.99$

Feature 1:

$H(T) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.811$

$H(F) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.721$

$H(F_1) = \frac{4}{9}\cdot 0.811 + \frac{5}{9}\times 0.721 = 0.76$

$IG(F_1) = 0.23$

Feature 2:

$H(T) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.97$

$H(F) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$H(F_2) = 0.971 \times \frac{5}{9} + 1\times\frac{4}{9} = 0.982$

$IG(F_2) = 0.99 - 0.924 = 0.008$

3.2:

H(label) = 0.99

threhold = 2.5

$H(S) = 0$

$H(G) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.953$

$H(2.5) = 0.953 \times \frac{8}{9} = 0.847$

$IG(F=2.5) = 0.143$

threhold = 3.5

$H(S) = 1$

$H(G) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7} = 0.985$

$IG(3.5) = 0.003$

threhold = 4.5

$H(S) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.917$

$H(G) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.917$

$H(4.5) = 0.917\cdot(\frac{3}{9} + \frac{6}{9}) = 0.917$

$IG(4.5) = 0.073$

2.5:

from the equation ② & ①, we have:

$(x_1+x_2)\log_2\frac{y_1+y_2}{x_1+x_2} \geq x_1\log_2\frac{y_1}{x_1} + x_2\log_2\frac{y_2}{x_2}$

∵ since $n_m^+ = n_{ms}^+ + n_{mu}^+$

$\therefore n_{mu}^+\log_2\frac{n_{mu}^+}{n_{mu}} + n_{ms}^+\log_2\frac{n_{ms}^+}{n_{ms}} \geq n_m^+\log_2\frac{n_m^+}{n_m}$ ①

$n_{mu}^-\log_2\frac{n_{mu}^-}{n_{mu}} + n_{ms}^-\log_2\frac{n_{ms}^-}{n_{ms}} \geq n_m^-\log_2\frac{n_m^-}{n_m}$ ②

When ① + ②, we will get:

$n_{mu}^+\log_2\frac{n_{mu}^+}{n_{mu}} + n_{ms}^+\log_2\frac{n_{ms}^+}{n_{ms}} + n_{mu}^-\log_2\frac{n_{mu}^-}{n_{mu}} + n_{ms}^-\log_2\frac{n_{ms}^-}{n_{ms}}$

$\geq n_m^+\log_2\frac{n_m^+}{n_m} + n_m^-\log_2\frac{n_m^-}{n_m}$

∴ by divide n on both side and made some slight change:

$-\frac{n_{mu}^+}{n}\log_2\frac{n_{mu}}{n_{mu}^+} + (-\frac{n_{ms}^+}{n}\log_2\frac{n_{ms}}{n_{ms}^+}) + (-\frac{n_{mu}^-}{n}\log_2\frac{n_{mu}}{n_{mu}^-}) + (-\frac{n_{ms}^-}{n}\log_2\frac{n_{ms}}{n_{ms}^-}) \geq -\frac{n_m^+}{n}\log_2\frac{n_m}{n_m^+} + (-\frac{n_m^-}{n}\log_2\frac{n_m}{n_m^-})$

$\therefore -\frac{n_{mu}^+}{n}\log_2\frac{n_{mu}}{n_{mu}^+} - \frac{n_{mu}^-}{n}\log_2\frac{n_{mu}}{n_{mu}^-} - \frac{n_{ms}^+}{n}\log_2\frac{n_{ms}}{n_{ms}^+} - \frac{n_{ms}^-}{n}\log_2\frac{n_{ms}}{n_{ms}^-} \geq -\frac{n_m^+}{n}\log_2\frac{n_m}{n_m^+} - \frac{n_m^-}{n}\log_2\frac{n_m}{n_m^-}$

$-\frac{n_{mu}}{n}(\frac{n_{mu}^+}{n_{mu}}\log_2\frac{n_{mu}}{n_{mu}^+} + \frac{n_{mu}^-}{n_{mu}}\log_2\frac{n_{mu}}{n_{mu}^-}) - \frac{n_{ms}}{n}(\frac{n_{ms}^+}{n_{ms}}\log_2\frac{n_{ms}}{n_{ms}^+} + \frac{n_{ms}^-}{n_{ms}}\log_2\frac{n_{ms}}{n_{ms}^-}) \geq -\frac{n_m}{n}(\frac{n_m^+}{n_m}\log_2\frac{n_m}{n_m^+} + \frac{n_m^-}{n_m}\log_2\frac{n_m}{n_m^-})$

∴

There Fore:

$\therefore H(S) - \sum_{j=1}^{a}\frac{|S_j|}{|S|}\times H(S_j) \geq H(S) - \sum_{j=1}^{a}\frac{|S_{ij}|}{|S|}\cdot H(S_{ij})$

$\therefore IG(SSF) \geq IG(SS)$

threhold = 5.5

$H(S) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5} = 0.971$

$H(G) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$H(C|S) = \frac{5}{9} \cdot 0.971 + \frac{4}{9} = 0.983$

$IG(\text{threhold}=5.5) = 0.007$

threhold = 6.5

$H(S) = 1$

$H(G) = -\frac{1}{3}\log_2\frac{1}{2} - \frac{2}{3}\log_2\frac{2}{3} = 0.917$

$IG(6.5) = 0.99 - 1 \times \frac{6}{9} - 0.917 \times \frac{3}{9} = 0.017$

threhold = 7.5

$H(S) = 1$

$H(G) = 0$

$IG = 0.99 - \frac{8}{9} = 0.102$

∴ therefor, when threhold = 2.5, the IG is highest

3.3.

Feature No: 1 will be my root node

Since it has hightest IG

3.4