Assignment 4

Ch 3 | ex 2, 3, 5

(2) a) Gini index → $Gini(t) = 1 - \sum_{i=1}^{c-1} [p(i|t)]^2$

$Gini(class) = 1 - \left[\left(\frac{10}{20}\right)^2 + \left(\frac{10}{20}\right)^2\right] = 0.5$

b) $Gini(ID) = 0$

c) $Gini(Gender) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{4}{10}\right)^2\right] = \boxed{0.48}$

for Male of Female

d) Gini (Car Type) = ?

=

for sports cars

Gini(family car) = $1 - \left[\left(\frac{1}{4}\right)^2 + \left(\frac{3}{4}\right)^2\right] = \boxed{0.375}$

Gini (Sports car) = $1 - 1 = \boxed{0}$

↓

it's $\left(\frac{8}{8}\right)^2 = 1$

Gini (Luxury car) = $1 - \left[\left(\frac{1}{8}\right)^2 + \left(\frac{7}{8}\right)^2\right] = \boxed{0.21875}$

$\Rightarrow$ Gini (Car Type) = $\frac{4}{20} \times 0.375 + \frac{8}{20} \times 0.21875 +$

$+ \frac{8}{20} \times 0 = \boxed{0.163}$

e) Gini (Shirt Size) = ?

Gini(Small) = $1 - \left[\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] = 0.48$

Gini (Medium) = $1 - \left[\left(\frac{3}{7}\right)^2 + \left(\frac{4}{7}\right)^2\right] = 0.49$

Gini (Large) = $1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$

Gini (XL) = $1 - \left[\left(\frac{2}{4}\right)^2 + \left(\frac{2}{4}\right)^2\right] = 0.5$

Gini (Shirt Size) = $\frac{5}{20} \times 0.48 + \frac{7}{20} \times 0.49 + \frac{4}{20} \times 0.5 +$

$+ \frac{4}{20} \times 0.5 = \underline{0.4915}$

f) Car type because it has the lowest Gini index

g) Because it has no predictive power. Every new customer has a new ID.

3) a) $H(TC) = -\left[\frac{4}{9}\log_2\frac{4}{9} + \frac{5}{9}\log_2\frac{5}{9}\right] = \boxed{0.99}$

b) $H(TC|a_1) = -\left[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right] = \boxed{0.81}$

$H(TC|a_2) = -\left[\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5}\right] = \boxed{0.72}$

$\Rightarrow H(TC|a_1) = \frac{4}{9}\times 0.811 + \frac{5}{9}\times 0.72 = \boxed{0.76}$

$\Rightarrow I_G = 0.99 - 0.76 = \boxed{0.23}$

$H(TC|a_2) = \frac{4}{9}\times\left[-\left(\frac{2}{4}\log_2\frac{2}{4} + \frac{2}{4}\log_2\frac{2}{4}\right)\right] +$

$+ \frac{5}{9}\times\left[-\left(\frac{3}{5}\log_2\frac{3}{5} + \frac{2}{5}\log_2\frac{2}{5}\right)\right] = \boxed{0.98}$

$\Rightarrow I_G = 0.99 - 0.98 = \boxed{0.01}$

c)

a3	Class label	Split point	Entropy	IG
1.0	+	2.0	0.848	0.142
3.0	−	3.5	0.998	0.01
4.0	+	4.5	0.918	0.07
5.0	−	5.5	0.994	0.01
5.0	−			

6.0	+	6.5	0.9768	0.02
7.0	+	7.5	0.889	0.10
7.0	−			
8.0	−			

↳ don't need to include it in the entropy

d) it's a_1

e) Misclassification error rate

$a_1 = \frac{2}{9}$ $a_2 = \frac{5}{9}$

$\Rightarrow a_1$ is the best split attribute

f) $Gini(a_{1+}) = 1 - \left[\left(\frac{3}{4}\right)^2 + \left(\frac{1}{4}\right)^2\right]$

$Gini(a_{1-}) = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{4}{5}\right)^2\right]$

$\Rightarrow Gini(a_1) = \frac{4}{9} \times Gini(a_{1+}) + \frac{5}{9} \times Gini(a_{1-}) = \boxed{0.344}$

$Gini(a_2) = \frac{4}{9} \times \left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2\right] +$

$+ \frac{5}{9} \times \left[1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2\right] = \boxed{0.489}$

a_1 is the best split attribute since $a_1 < a_2$

(5) a) Entropy

$$E(\text{class}) = -\left[\frac{4}{10}\log_2\frac{4}{10} + \frac{6}{10}\log_2\frac{6}{10}\right] = 0.97$$

$$E(A) = \frac{7}{10}\cdot\left[-\left(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7}\right)\right] + \frac{3}{10}\left[-\left(\frac{3}{3}\log_2\frac{3}{3} + \frac{0}{3}\log_2\frac{0}{3}\right)\right] = \boxed{0.1813}$$

$$E(B) = \frac{4}{10}\cdot\left[\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right] + \frac{6}{10}\left[\frac{1}{6}\log_2\frac{1}{6} + \frac{5}{6}\log_2\frac{5}{6}\right] = \boxed{0.2565}$$

$IG_A = 0.97 - 0.2103$

$IG_B = 0.97 - 0.2565$ → The decision tree induction will choose this attribute

6) $G = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$

$T = 7$, $F = 3$

For T: $+ = 4$, $- = 3$

For F: $+ = 0$, $- = 3$

$$G(T) = 1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2 = 0.493$$

$G(F) = 1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2 = 0$

$\Rightarrow G = \left(\frac{7}{10}\right) \times 0.493 + \left(\frac{3}{10}\right) \times 0 = 0.34$

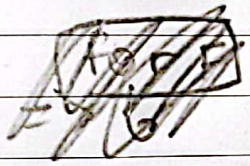
↳ Gini index for A

$\Rightarrow$ ~~T=4~~, F=6
T=4

For T: +=3, -=1
For F: +=1, -=5

$\Rightarrow G(T) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$

$G(F) = 1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2 = 0.286$



$\rightarrow G = \left(\frac{4}{10}\right) \times 0.375 + \left(\frac{6}{10}\right) \times 0.286 = 0.32$

↳ Gini index for B

→ Gini for B is less => it will be chosen to split the node

c) Yes, it's possible. The measures have similar range and monotonous behaviour in their information gain. The results of the Δ will favour different attributes.

Ch.4, ex.18

a) There are equal number of positive and negative records in the data and the classifier predicts every test record as positive $\Rightarrow$ error rate is 0.5 (half of the samples are misclassified).

b) Samples $\# = n$

$\frac{n}{2}$ can be misclassified as negative with probability 0.8

$\frac{n}{2}$ can be missed as positive with prob. 0.2

$$\Rightarrow \text{error rate} = \frac{0.8 \times \frac{n}{2} + 0.2 \times \frac{n}{2}}{n} = \boxed{0.5}$$

c) if $\frac{2}{3}$ of the data is positive and $\frac{1}{3}$ is negative and we assume that everything is positive

$\Rightarrow$ the error rate is $\boxed{0.333...}$

d) error rate $= \frac{\frac{2}{3} \times \frac{2}{3}\pi + \frac{1}{3} \times \frac{1}{3}\pi}{\pi} =$

$= \boxed{0.55...}$