Solution to Sample Tasks

Task 1

Entropy before split,

$x = 5,\ y = 5$

$$H = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1.$$

Using A to split.

$A = 1$

$x = 3,\ y = 0$

$$H_{A_1} = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$A = 2$

$x = 1,\ y = 3$

$$H_{A_2} = -\frac{1}{4}\log_2\frac{1}{4} - \frac{3}{4}\log_2\frac{3}{4} = 0.8113$$

$A = 3$

$x = 2,\ y = 2$

$$H_{A_3} = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9183$$

$$I_A = H - \frac{3}{10} H_{A_1} - \frac{4}{10} H_{A_2} - \frac{3}{10} H_{A_3}$$

$$= 0.4$$

Using B to split

$B = 1$

$x = 1, y = 3$

$$H_{B_1} = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.8113$$

$B = 2$

$x = 3, y = 1$

$$H_{B_2} = -\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} = 0.8113$$

$B = 3$

$x = 1, y = 1$

$$H_{B_3} = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1$$

$$I_B = H - \frac{4}{10} H_{B_1} - \frac{4}{10} H_{B_2} - \frac{2}{10} H_{B_3}$$

$$= 0.15096$$

Using C to split

$C=1$

$x:2 \quad y:3$

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

$$= 0.9709$$

$C=2$

$x:2 \quad y:2$

$$H_{C_2} = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$C=3$

$x:1 \quad y:0$

$$H_{C_3} = -\frac{1}{1}\log_2\frac{1}{1} - \frac{0}{1}\log_2\frac{0}{1} = 0$$

$$I_C = H - \frac{5}{10}H_{C_1} - \frac{4}{10}H_{C_2} - \frac{1}{10}H_{C_3}$$

$$= 0.11455//$$

So A is the best attribute.

Task 2

Before split,

$x:5,\ y:5$

$$H = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}$$

Consider A, 15

$A<15 \quad x:0,\ y:2$

$$H_{A<15} = -\frac{0}{2}\log_2\frac{0}{2} - \frac{2}{2}\log_2\frac{2}{2} = 0$$

$A\geq 15 \quad x:5,\ y:3$

$$H_{A\geq 15} = -\frac{5}{8}\log_2\frac{5}{8} - \frac{3}{8}\log_2\frac{3}{8} = 0.9544$$

$$I_{A,15} = 1 - \frac{2}{10}(0) - \frac{8}{10}(0.9544) = 0.2365$$

Consider $A, 20$

$A < 20$

$x : 0, y = 3$

$$H_{A<20} = -\frac{0}{3}\log_2\frac{0}{3} - \frac{3}{3}\log_2\frac{3}{3} = 0$$

$A \geq 20$

$x : 5, y = 2$

$$H_{A\geq 20} = -\frac{5}{7}\log_2\frac{5}{7} - \frac{2}{7}\log_2\frac{2}{7} = 0.8631$$

$$I_{A,20} = 1 - \frac{3}{10}(0) - \frac{7}{10}(0.8631) = 0.3958$$

Consider $A, 25$

$A < 25$

$x : 2, y = 5$

$$H_{A<25} = -\frac{2}{7}\log_2\frac{2}{7} - \frac{5}{7}\log_2\frac{5}{7} = 0.8631$$

$A \geq 25$

$x : 3, y = 0$

$$H_{A\geq 25} = -\frac{3}{3}\log_2\frac{3}{3} - \frac{0}{3}\log_2\frac{0}{3} = 0$$

$$I_{A,25} = 1 - \frac{7}{10}(0.8631) - \frac{3}{10}(0) = 0.3958$$

Consider B, 15

$B < 15$

$x = 2, y = 0$

$$H_{B<15} = -\frac{2}{2}\log_2\frac{2}{2} - \frac{0}{2}\log_2\frac{0}{2} = 0$$

$B \geq 15$

$x = 3, y = 5$

$$H_{B\geq 15} = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8} = 0.9544$$

$$I_{B,15} = 1 - \frac{2}{10}(0) - \frac{8}{10}(0.9544) = 0.2365$$

Consider B, 20

$B < 20$

$x = 2, y = 0$

$$H_{B<20} = 0$$

$B \geq 20$

$x = 3, y = 5$

$$H_{B\geq 20} = 0.9544$$

$$I_{B,20} = 0.2365$$

Consider $B, 25$

$B < 25$ $\quad x: 5, y: 0$

$H_{B<25} = 0$

$B_{\geq 25} = x: 0 \quad y: 5$

$H_{B \geq 25} = 0$

$$IG_{B,25} = 1 - \frac{5}{10}(0) - \frac{5}{10}(0) = 1$$

B with the threshold of 25, has the highest possible Information Gain and so is the best attribute.

Task 3

Using Histogram:

No Binning: 128^{30}

With Binning: 32^{30}

Using Gaussians:

1-D Gaussians:

Per Gaussian: 2

In Total: 60

Mixed-D Gaussians (3):

Per Gaussian: 9

In Total: 270

30-D Gaussian.

In Theory:

μ: 30

Σ: 30^2

Total. 930

In Practice:

μ: 30

Σ: 465 [450 is also acceptable]

Total: 495

[480 is also acceptable]

Mix of 30-D Gaussian (3)

In Theory (Per Gaussian):

μ: 30

Σ: 30^2

w: 1

Total: 2793

In Practice (Per Gaussian):

μ: 30

Σ: 465 [or 450]

w: 1

Total: 1488 [or 1443]