Joshua Krachman
Quiz 6 Answers
Biomedical Data Science

1)

$$H(y) = -\sum_{i=1}^{n} P_i \log(P_i) =$$

$$-\left[\frac{6}{10}\log_2\frac{6}{10} + \frac{4}{10}\log_2\frac{4}{10}\right] = 0.97095$$

$$H(y) = 0.97095$$

2)

$$Information\ Gain\ IG = H_p - \left(\frac{C_1}{P}H_c^1 + \frac{C_2}{P}H_c^2\right)$$

$$x_1 = author, x_2 = long, x_3 = research, x_4 = grade, x_5 = lottery$$

$$IG_{x1} = 0.97095 - [\frac{6}{10} * -1[(\frac{3}{6}\log_2\frac{3}{6} + \frac{3}{6}\log_2\frac{3}{6})] + \frac{4}{10} * -1[(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4})]]$$

$$IG_{x1} = 0.04644$$

$$IG_{x2} = 0.97095 - [\frac{5}{10} * -1[(\frac{5}{5}\log_2\frac{5}{5} + \frac{0}{5}\log_2\frac{0}{5})] + \frac{5}{10} * -1[(\frac{1}{5}\log_2\frac{1}{5} + \frac{4}{5}\log_2\frac{4}{5})]]$$

$$IG_{x2} = 0.60999$$

$$IG_{x3} = 0.97095 - [\frac{7}{10} * -1[(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7})] + \frac{3}{10} * -1[(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3})]]$$

$$IG_{x3} = 0.00580$$

$$IG_{x4} = 0.97095 - [\frac{7}{10} * -1[(\frac{5}{7}\log_2\frac{5}{7} + \frac{2}{7}\log_2\frac{2}{7})] + \frac{3}{10} * -1[(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3})]]$$

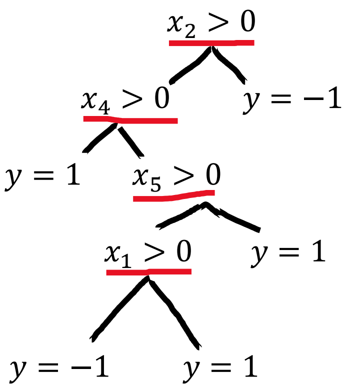
$$IG_{x4} = 0.09128$$

$$IG_{x5} = 0.97095 - [\frac{3}{10} * -1[(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3})] + \frac{7}{10} * -1[(\frac{4}{7}\log_2\frac{4}{7} + \frac{3}{7}\log_2\frac{3}{7})]]$$

$$IG_{x5} = 0.00580$$

You should split on x_2 for the root node because it has largest Information Gain.

3) *TREE: If the statement is true, go to the right. If false, go to the left.*


$x_2 > 0$
$x_4 > 0$
$y = -1$
$y = 1$
$x_5 > 0$
$x_1 > 0$
$y = 1$
$y = -1$
$y = 1$