

Question 2

Part a

- 1) 4.0 GPA - P, the data match the decision tree
- 2) 3.9 GPA - P, the data match the decision tree
- 3) 3.9 GPA - P, the data match the decision tree
- 4) 3.8 GPA - yes publications - P, the data match the decision tree
- 5) 3.6 GPA - no publications - rank 2 university - P, the data match the decision tree
- 6) 3.6 GPA - yes publications - P, the data match the decision tree
- 7) 3.4 GPA - no publications - rank 3 university - N, the data match the decision tree
- 8) GPA 3.4 - No publication - Rank 1 University - N, data match the tree
- 9) GPA 3.2 - N, data match the tree
- 10) GPA 3.1 - N, data match the tree
- 11) GPA 3.1 - N, data match the tree
- 12) GPA 3.0 - N, data match the tree

Part b

$$E(S) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} = 1 \quad (1)$$

For GPA, the information gained is:

$$E(GPA = 4.0) = -1 \log_2 1 - 0 \log_2 0 = 0 \quad (2)$$

$$E(GPA = 3.6) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.9710 \quad (3)$$

$$E(GPA = 3.3) = -0 \log_2 0 - 1 \log_2 1 = 0 \quad (4)$$

$$I(GPA) = \frac{1}{4} * 0 + \frac{5}{12} * 0.9710 + \frac{1}{3} * 0 = 0.4046 \quad (5)$$

$$Gain(GPA) = E(S) - I(GPA) = 1 - 0.4046 = 0.5954 \quad (6)$$

For university rank, the information gained is:

$$E(rank = 1) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = 0.9710 \quad (7)$$

$$E(rank = 2) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3} = 0.9183 \quad (8)$$

$$E(rank = 3) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = 0.8113 \quad (9)$$

$$I(rank) = \frac{5}{12} * 0.9710 + \frac{1}{4} * 0.9183 + \frac{1}{3} * 0.8113 = 0.9046 \quad (10)$$

$$Gain(rank) = E(S) - I(rank) = 1 - 0.9046 = 0.0954 \quad (11)$$

For whether the student has publications, the information gained is:

$$E(Published = Yes) = -\frac{3}{5} \log_2 \frac{3}{5} - \frac{2}{5} \log_2 \frac{2}{5} = .9710 \quad (12)$$

$$E(Published = No) = -\frac{3}{7} \log_2 \frac{3}{7} - \frac{4}{7} \log_2 \frac{4}{7} = .9852 \quad (13)$$

$$I(Published) = \frac{5}{12}(.9710) + \frac{7}{12}(.9852) = 0.9792 \quad (14)$$

$$Gain(Published) = E(S) - I(Published) = 1 - 0.9792 = 0.0208 \quad (15)$$

For the quality of the student's recommendations, the information gained is:

$$E(Recommendations = Good) = -\frac{5}{8} \log_2 \frac{5}{8} - \frac{3}{8} \log_2 \frac{3}{8} = .9544 \quad (16)$$

$$E(Recommendations = Normal) = -\frac{1}{4} \log_2 \frac{1}{4} - \frac{3}{4} \log_2 \frac{3}{4} = .8113 \quad (17)$$

$$I(Recommendations) = \frac{8}{12}(.9544) + \frac{4}{12}(.8113) = 0.9067 \quad (18)$$

$$Gain(Recommendations) = E(S) - I(Recommendations) = 1 - 0.9067 = 0.0933 \quad (19)$$

For the root of the decision tree, the best attribute to use is GPA, since it has the highest information gain of all the attributes.