Homework 1

Question 1: Decision Tree Induction (10 points)

Consider the training set given below for determining whether a loan application should be approved or rejected. Draw the full decision tree obtained using entropy as the impurity measure. Show your steps clearly (i.e., the computation of information gain for every candidate attribute must be shown). Compute the training error of the decision tree.

Long-Term Debt	Unemployed	Credit Rating	Down Payment < 20%	Class
No	No	Good	Yes	Approve
No	No	Bad	No	Approve
No	No	Bad	Yes	Approve
No	No	Bad	No	Approve
Yes	No	Good	No	Approve
No	Yes	Good	Yes	Reject
Yes	No	Bad	No	Reject
Yes	No	Bad	Yes	Reject
Yes	No	Bad	Yes	Reject
Yes	Yes	Bad	No	Reject

Total entropy:

$$H(x) = -\left(\frac{5}{10}\log_2\frac{5}{10}\right) - \left(\frac{5}{10}\log_2\frac{5}{10}\right) = .5 \times -1 - (.5 \times -1) = 1$$

Long-term debt: 5 no - 4 accept, 1 reject; 5 yes - 1 accept, 4 reject

$$H(No) = -\left(\frac{4}{5}\log_2\frac{4}{5}\right) - \left(\frac{1}{5}\log_2\frac{1}{5}\right) = -(.8 \times -0.3219) - (0.2 \times -2.3219)$$
$$= 0.7214$$

$$H(yes) = -\left(\frac{1}{5}\log_2\frac{1}{5}\right) - \left(\frac{4}{5}\log_2\frac{4}{5}\right) = 0.7214$$

$$H(debt) = 0.7214$$

$$Gain = 1 - 0.7214 = 0.2781$$

Unemployed: 8 N - 5A, 3R; 2Y - 0A, 2R

$$H(N) = -\left(\frac{5}{8}\log_2\frac{5}{8}\right) - \left(\frac{3}{8}\log_2\frac{3}{8}\right) = .9544$$

$$H(Y) = 0$$

$$H(unemployed) = \frac{8}{10}(.9544) + 0 = 0.7635$$

$$Gain = 1 - 0.7635 = 0.2365$$

Credit: 3 Good: 2A, 1R; 7 Bad: 3A, 4R

$$H(G) = -\left(\frac{2}{3}\log_2\frac{2}{3}\right) - \left(\frac{1}{3}\log_2\frac{1}{3}\right) = 0.9183$$

Credit: 3 Good: 2A, 1R; 7 Bad: 3A, 4R

$$H(G) = -\left(\frac{2}{3}\log_2\frac{2}{3}\right) - \left(\frac{1}{3}\log_2\frac{1}{3}\right) = 0.9183$$

$$H(B) = -\left(\frac{3}{7}\log_2\frac{3}{7}\right) - \left(\frac{4}{7}\log_2\frac{4}{7}\right) = 0.9852$$

$$H(credit) = \frac{3}{10}(.9183) + \frac{7}{10}(0.9852) = 0.9651$$
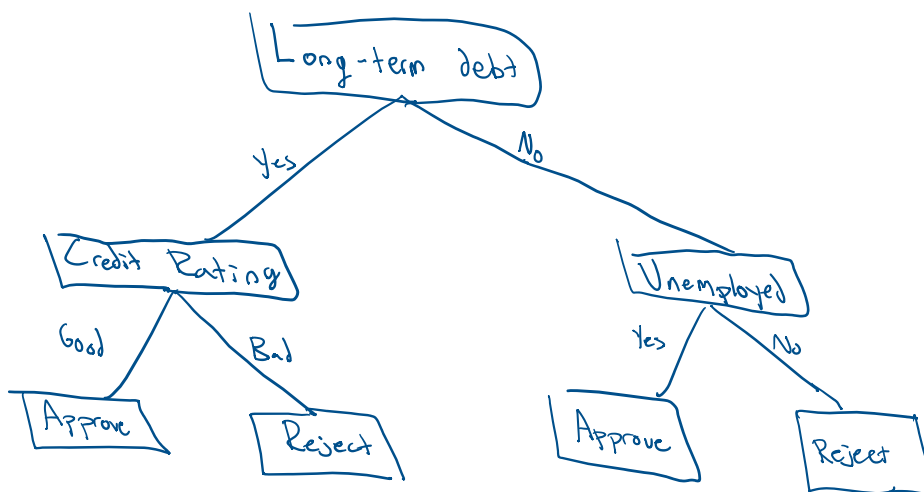
$$Gain = 1 - 0.9651 = 0.0349$$

Down Payment: 5Y: 3A, 2R; 5N: 2A, 3R

$$H(Y) = -\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \left(\frac{2}{5}\log_2\frac{2}{5}\right) = 0.9710$$

$$H(N) = -\left(\frac{2}{5}\log_2\frac{2}{5}\right) - \left(\frac{3}{5}\log_2\frac{3}{5}\right) = 0.9710$$

$$H(Down) = 0.9710$$

$$Gain = 1 - 0.9710 = 0.029$$

Long-term debt
- Yes → Credit Rating
  - Good → Approve
  - Bad → Reject
- No → Unemployed
  - Yes → Approve
  - No → Reject

Training error = 0/10 = 0