1. Decision Trees (5 points)–Complete the last node of the decision tree that we had done in class. Clearly show all you work in computing the intermediate results and show the formulas used. Draw out the final tree showing the decisions resulting at each node in the tree.

Two training samples to work with: {2,3}

Two remaining attributes: 1. Collateral; 2. Debt.

Total information content is:

$$I[Loans] = \sum_{i=1}^{3} -\frac{c_i}{C} \cdot log_2\left(\frac{c_i}{C}\right) = -\frac{6}{14} \cdot log_2\left(\frac{6}{14}\right) - \frac{3}{14} \cdot log_2\left(\frac{3}{14}\right) - \frac{5}{14} \cdot log_2\left(\frac{5}{14}\right) = 1.531$$

Formula for expected information and the corresponding gain:

$$E[Attribute] = \sum_{i=1}^{N_{classes}} \frac{c_i}{C} \cdot I\left(\frac{c_i}{C}\right)$$

At the last node the Expected Information from Collateral is:

– Collateral has one level:

- None - 2 instances out of 2 (samples 2,3)

$$E[Collateral] = \frac{c_{none}}{C} \cdot I(c_{none}) + \frac{c_{adequate}}{C} \cdot I(c_{adequate}) = \frac{2}{2} \cdot I(c_{none}) + \frac{0}{2} \cdot I(c_{adequate})$$

– And now:

$$I(c_{none}) = -p_{low\_risk} \cdot log_2(p_{low\_risk}) - p_{med\_risk} \cdot log_2(p_{med\_risk}) - p_{high\_risk} \cdot log_2(p_{high\_risk})$$

– where here p_{low_risk} is the frequency of no collateral training samples that are low risk.

Looking at training samples we see 0 no_collateral training samples are low risk.

– where here p_{med_risk} is the frequency of no collateral training samples that are med risk.

Looking at training samples we see 1 no_collateral training samples are med risk.

– where here p_{high_risk} is the frequency of no collateral training samples that are high risk.

Looking at training samples we see 1 no_collateral training samples are high risk.

$$I(c_{none}) = -p_{low_{risk}} \cdot log_2(p_{low_{risk}}) - p_{med_{risk}} \cdot log_2(p_{med_{risk}}) - p_{high_{risk}} \cdot log_2(p_{high_{risk}})$$
$$= -0 \cdot log_2(0) - \frac{1}{2} \cdot log_2\left(\frac{1}{2}\right) - \frac{1}{2} \cdot log_2\left(\frac{1}{2}\right) = 1$$

$$I(c_{adequate}) = -p_{low_{risk}} \cdot log_2(p_{low_{risk}}) - p_{med_{risk}} \cdot log_2(p_{med_{risk}}) - p_{high_{risk}} \cdot log_2(p_{high_{risk}})$$
$$= -0 \cdot log_2(0) - 0 \cdot log_2(0) - 0 \cdot log_2(0) = 0$$

At the last node the Expected Information from Collateral then:

$$E[Collateral] = \frac{c_{none}}{C} \cdot I(c_{none}) + \frac{c_{adequate}}{C} \cdot I\left(c_{adequate}\right) = \frac{2}{2} \cdot I(c_{none}) + \frac{0}{2} \cdot I\left(c_{adequate}\right)$$
$$= \frac{2}{2} \cdot 1 + \frac{0}{2} \cdot 0 = 1$$

Therefore, the information **gain** at this node from using **collateral** is:

$$gain(collateral) = I[training\_set] - E[Collateral] = 1.531 - 1 = 0.531$$

At the last node the Expected Information from Debt is:

– Debt has two levels:

• Low debt – 1 instance out of 2 (samples 3)

• High debt – 1 instance out of 2 (samples 2)

$$E[Debt] = \frac{c_{low\_debt}}{C} \cdot I\left(c_{low\_debt}\right) + \frac{c_{high\_debt}}{C} \cdot I\left(c_{high\_debt}\right) = \frac{1}{2} \cdot I\left(c_{low\_debt}\right) + \frac{1}{2} \cdot I\left(c_{high\_debt}\right)$$

– And now:

$$I\left(c_{low\_debt}\right) = -p_{low\_risk} \cdot log_2\left(p_{low\_risk}\right) - p_{med\_risk} \cdot log_2\left(p_{med\_risk}\right) - p_{high\_risk} \cdot log_2\left(p_{high\_risk}\right)$$

– where here p_{low_risk} is the frequency of low debt training samples that are low risk.

Looking at training samples we see 0 low debt training samples are low risk.

– where here p_{med_risk} is the frequency of low debt training samples that are med risk.

Looking at training samples we see 1 low debt training samples are med risk.

– where here p_{high_risk} is the frequency of low debt training samples that are high risk.

Looking at training samples we see 0 low debt training samples are high risk.

$$I\left(c_{low\_debt}\right) = -p_{low_{risk}} \cdot log_2\left(p_{low_{risk}}\right) - p_{med_{risk}} \cdot log_2\left(p_{med_{risk}}\right) - p_{high_{risk}} \cdot log_2\left(p_{high_{risk}}\right)$$
$$= -0 \cdot log_2(0) - \frac{1}{2} \cdot log_2\left(\frac{1}{2}\right) - 0 \cdot log_2(0) = 0.5$$

– And also

$$I\left(c_{high\_debt}\right) = -p_{low\_risk} \cdot log_2\left(p_{low\_risk}\right) - p_{med\_risk} \cdot log_2\left(p_{med\_risk}\right) - p_{high\_risk} \cdot log_2\left(p_{high\_risk}\right)$$

– where here p_{low_risk} is the frequency of high debt training samples that are low risk.

Looking at training samples we see 0 high debt training samples are low risk.

– where here p_{med_risk} is the frequency of high debt training samples that are med risk.

Looking at training samples we see 0 high debt training samples are med risk.

– where here p_{high_risk} is the frequency of high debt training samples that are high risk.

Looking at training samples we see 1 high debt training samples are high risk.

$$(c_{high\_debt}) = -p_{low_{risk}} \cdot log_2(p_{low_{risk}}) - p_{med_{risk}} \cdot log_2(p_{med_{risk}}) - p_{high_{risk}} \cdot log_2(p_{high_{risk}})$$
$$= -0 \cdot log_2(0) - 0 \cdot log_2(0) - \frac{1}{2} \cdot log_2\left(\frac{1}{2}\right) = 0.5$$

At the last node the Expected Information from Debt then:

$$E[Debt] = \frac{c_{low\_debt}}{C} \cdot I(c_{low_{debt}}) + \frac{c_{high_{debt}}}{C} \cdot I(c_{high_{debt}}) = \frac{1}{2} \cdot I(c_{low_{debt}}) + \frac{1}{2} \cdot I(c_{high_{debt}})$$
$$= \frac{1}{2} \cdot 0.5 + \frac{1}{2} \cdot 0.5 = 0.5$$

Therefore the information **gain** at this node from using **Debt** is:

$$gain(debt\_level) = I[training\_set] - E[debt_{level}] = 1.531 - 0.5 = 1.031$$

Summary of Information Gains:

- If first attribute is Collateral:

– Gain(Collateral) = I[C] – E[Collateral]

– Gain(Collateral) = 1.531 – 1 = 0.531

- If first attribute is Debt:

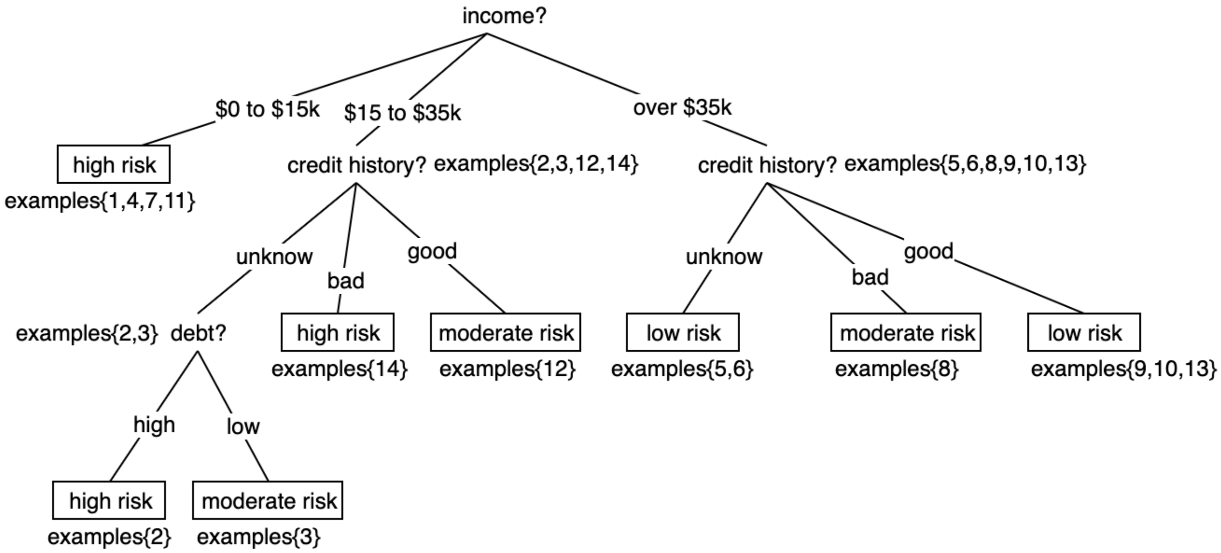
– Gain(Debt) = I[C] – E[Debt]

– Gain(Debt) = 1.531 – 0.5= 1.031

- **So the best attribute for the last node is Debt**

– Gain(Debt) = 1.531 – 0.5= 1.031

– This makes a decision tree that looks like:


income?
$0 to $15k
$15 to $35k
over $35k
high risk
examples{1,4,7,11}
credit history? examples{2,3,12,14}
credit history? examples{5,6,8,9,10,13}
unknow
bad
good
unknow
bad
good
examples{2,3} debt?
high risk
examples{14}
moderate risk
examples{12}
low risk
examples{5,6}
moderate risk
examples{8}
low risk
examples{9,10,13}
high
low
high risk
examples{2}
moderate risk
examples{3}