Decision Tree for Buys Computer Problem

Your Name

1 Dataset

The dataset contains information about customers, and we aim to predict whether they buy a computer based on attributes like Age, Income, Student status, and Credit Rating.

RID	Age	Income	Student	Credit Rating	Buys Computer?
1	Youth	High	No	Fair	No
2	Youth	High	No	Excellent	No
3	Middle Aged	High	No	Fair	Yes
4	Senior	Medium	No	Fair	Yes
5	Senior	Low	Yes	Fair	Yes
6	Senior	Low	Yes	Excellent	No
7	Middle Aged	Low	Yes	Excellent	Yes
8	Youth	Medium	No	Fair	No
9	Youth	Low	Yes	Fair	Yes
10	Senior	Medium	Yes	Fair	Yes
11	Youth	Medium	Yes	Excellent	Yes
12	Middle Aged	Medium	No	Excellent	Yes
13	Middle Aged	High	Yes	Fair	Yes
14	Senior	Medium	No	Excellent	No

Table 1: Class-labeled Training Tuples from the AllElectronics Customer Database

2 Entropy of the Target Variable

The entropy of the target variable (Buys Computer) is calculated as:

$$H(S) = -p_{yes} \log_2(p_{yes}) - p_{no} \log_2(p_{no})$$

Where:

$$p_{yes} = \frac{9}{14}, \quad p_{no} = \frac{5}{14}$$

$$H(S) = -\left(\frac{9}{14} \log_2 \frac{9}{14}\right) - \left(\frac{5}{14} \log_2 \frac{5}{14}\right) = 0.94$$

3 Entropy and Information Gain Calculations

We now calculate the entropy for each attribute and the corresponding information gain.

3.1 Entropy for Age

- **Youth**: 5 instances (2 yes, 3 no)

$$H(Youth) = -\left(\frac{2}{5}\log_2\frac{2}{5}\right) - \left(\frac{3}{5}\log_2\frac{3}{5}\right) = 0.97$$

- **Middle Aged**: 4 instances (4 yes, 0 no)

$$H(MiddleAged) = -\left(\frac{4}{4}\log_2\frac{4}{4}\right) = 0.0$$

- **Senior**: 5 instances (3 yes, 2 no)

$$H(Senior) = -\left(\frac{3}{5}\log_2\frac{3}{5}\right) - \left(\frac{2}{5}\log_2\frac{2}{5}\right) = 0.97$$

The weighted entropy for Age is:

$$H(Age) = \frac{5}{14}\cdot 0.97 + \frac{4}{14}\cdot 0 + \frac{5}{14}\cdot 0.97 = 0.693$$

Information Gain for Age:

$$\text{Gain}(S, Age) = 0.94 - 0.693 = 0.247$$

3.2 Entropy for Income

- **High**: 4 instances (2 yes, 2 no)

$$H(High) = -\left(\frac{2}{4}\log_2\frac{2}{4}\right) - \left(\frac{2}{4}\log_2\frac{2}{4}\right) = 1.0$$

- **Medium**: 4 instances (2 yes, 2 no)

$$H(Medium) = -\left(\frac{2}{4}\log_2\frac{2}{4}\right) - \left(\frac{2}{4}\log_2\frac{2}{4}\right) = 1.0$$

- **Low**: 6 instances (5 yes, 1 no)

$$H(Low) = -\left(\frac{5}{6}\log_2\frac{5}{6}\right) - \left(\frac{1}{6}\log_2\frac{1}{6}\right) = 0.65$$

The weighted entropy for Income is:

$$H(Income) = \frac{4}{14} \cdot 1.0 + \frac{4}{14} \cdot 1.0 + \frac{6}{14} \cdot 0.65 = 0.857$$

Information Gain for Income:

$$\text{Gain}(S, Income) = 0.94 - 0.857 = 0.083$$

3.3 Entropy for Student

- **Yes**: 6 instances (5 yes, 1 no)

$$H(Yes) = -\left(\frac{5}{6} \log_2 \frac{5}{6}\right) - \left(\frac{1}{6} \log_2 \frac{1}{6}\right) = 0.65$$

- **No**: 8 instances (4 yes, 4 no)

$$H(No) = -\left(\frac{4}{8} \log_2 \frac{4}{8}\right) - \left(\frac{4}{8} \log_2 \frac{4}{8}\right) = 1.0$$

The weighted entropy for Student is:

$$H(Student) = \frac{6}{14} \cdot 0.65 + \frac{8}{14} \cdot 1.0 = 0.836$$

Information Gain for Student:

$$\text{Gain}(S, Student) = 0.94 - 0.836 = 0.104$$

3.4 Entropy for Credit Rating

- **Fair**: 8 instances (6 yes, 2 no)

$$H(Fair) = -\left(\frac{6}{8} \log_2 \frac{6}{8}\right) - \left(\frac{2}{8} \log_2 \frac{2}{8}\right) = 0.81$$

- **Excellent**: 6 instances (3 yes, 3 no)

$$H(Excellent) = -\left(\frac{3}{6} \log_2 \frac{3}{6}\right) - \left(\frac{3}{6} \log_2 \frac{3}{6}\right) = 1.0$$

The weighted entropy for Credit Rating is:

$$H(CreditRating) = \frac{8}{14} \cdot 0.81 + \frac{6}{14} \cdot 1.0 = 0.89$$

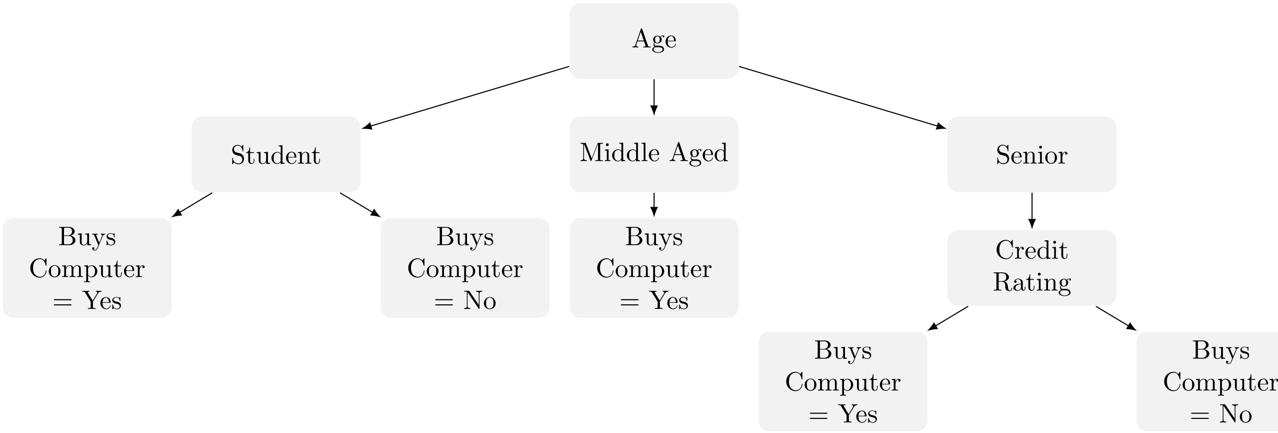
Information Gain for Credit Rating:

$$\text{Gain}(S, CreditRating) = 0.94 - 0.89 = 0.05$$

4 Decision Tree Diagram

- Overall entropy, $H(S) = 0.94$
- Entropy for Age: $H(Age) = 0.693$
- Information Gain for Age: $\text{Gain}(S, Age) = 0.247$
- Entropy for Income: $H(Income) = 0.857$
- Information Gain for Income: $\text{Gain}(S, Income) = 0.083$
- Entropy for Student: $H(Student) = 0.836$
- Information Gain for Student: $\text{Gain}(S, Student) = 0.104$
- Entropy for Credit Rating: $H(CreditRating) = 0.89$
- Information Gain for Credit Rating: $\text{Gain}(S, CreditRating) = 0.05$

Below is the diagram of the decision tree, where the root node is based on the attribute with the highest information gain.


Age
Student
Middle Aged
Senior
Buys Computer = Yes
Buys Computer = No
Buys Computer = Yes
Credit Rating
Buys Computer = Yes
Buys Computer = No