Brandon Nguyen
COMPE510 - Fall 2025
827813045

Programming Assignment 6 - Decision Trees

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

Problem 1: Consider the training examples shown in the following table for a binary classification problem.

a) Compute the Gini index for the overall collection of training examples.

1a) Total: 20, $C0=10$, $C1=10 \rightarrow 10/20 = 0.5$

$Gini = 1 - \sum p_i^2 = 1 - (0.5^2 + 0.5^2) = 1 - (0.25 + 0.25) = \boxed{0.5}$

Overall Gini $= 0.5$

b) Compute the Gini index for the Customer ID attribute.

1b) Customer ID: 20 values

Gini(Customer ID) = sum((1/20*0)) = $\boxed{0}$

c) Compute the Gini index for the Gender attribute.

1c) Gender: Male 10 values, Female 10 values

Male: C0=6, C1=4

Female: C0=4, C1=6

Male: $1-(0.6^2+0.4^2)=0.48$

Female: $1-(0.4^2+0.6^2)=0.48$

Gini(Gender) = (10/20)0.48 + (10/20)0.48 = $\boxed{0.48}$

d) Compute the Gini index for the Car Type attribute using multiway split.

1d) Family: 4/20 → C0 = 1, C1=3

Sport: 8/20 → C0 = 8, C1=0 → pure

Luxury: 8/20 → C0 = 1, C1=7

Gini(Family) $=1-(0.25^2+0.75^2) = 0.375$

Gini(Sport) = pure = 0

Gini(Luxury) $=1-(0.125^2+0.875^2)=0.21875$

Gini(Car) = (4/20)0.375 + 0 + (8/20)0.21875 = $\boxed{0.1625}$

e) Compute the Gini index for the Shirt Size attribute using multiway split.

1e) Small: 5/20 → C0 = 3, C1 = 2
Medium: 7/20 → C0 = 3, C1 = 4
Large: 4/20 → C0 = 2, C1 = 2
Extra Large: 4/20 → C0 = 2, C1 = 2

$\text{Gini(Small)} = 1-(0.6^2+0.4^2) = 0.48$
$\text{Gini(Medium)} = 1-((3/7)^2+(4/7)^2) = 0.49$
$\text{Gini(Large)} = 1-(0.5^2+0.5^2) = 0.5$
$\text{Gini(Extra Large)} = 1-(0.5^2+0.5^2) = 0.5$

$\text{Gini(Shirt Size)} = (5/20 * 0.48) + (7/20 * 0.49) + (4/20 * 0.5) + (4/20 * 0.5) = \boxed{0.49}$

f) Which attribute is better, Gender, Car Type, or Shirt Size?

Gini (Gender) = 0.48
Gini (Car Type) = 0.1625
Gini (Shirt Size) = 0.49

Smaller Gini value → more pure splits → better

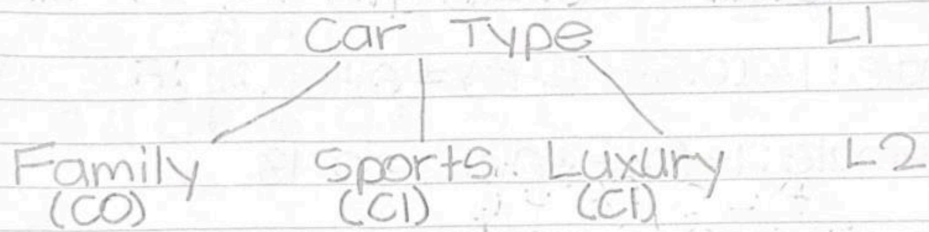
The best attribute between the three choices is Car Type, with the lowest Gini value.

g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

Customer ID should not be used as the attribute test condition even though it has the lowest Gini because it is only meant to tell us the total amount of things to take into account. The Gini value is 0 because it means nothing and overfits in terms of predictive value, therefore it is only an identifier for how many total data units there are to be taken into account from the chart.

h) Based on your results, construct a decision tree with 2 layers: the root node at the first layer, and the corresponding leaf nodes at the second layer.

1h)



i) Randomly generate a new data point (with valid attribute values), and use your 2-layer tree to predict its class label. How confident is your prediction? That is, what is the estimated probability that your prediction is correct, based on the training data reaching the corresponding leaf node?

1i) New Data Point: M / Luxury / Medium

C1 / Luxury = 7/8 (100) = 87.5%

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	−
4	F	F	4.0	+
5	F	T	7.0	−
6	F	T	3.0	−
7	F	F	8.0	−
8	T	F	7.0	+
9	F	T	5.0	−

Problem 2: Consider the training examples shown in the following table for a binary classification problem.

a) What is the entropy of this collection of training examples with respect to the positive class?

2a) $P(+) = 4/9 \quad P(-) = 5/9$

Entropy: $-\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right) = \boxed{0.991}$

b) What are the information gains of $a1$ and $a2$ relative to these training examples?

2b) a1:

$T \rightarrow P(+) = \frac{3}{4}, \; P(-) = \frac{1}{4}$

$F \rightarrow P(+) = \frac{1}{5}, \; P(-) = \frac{4}{5}$

$\text{Entropy}(T) = -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) = 0.811$

$\text{Entropy}(F) = -\frac{1}{5}\log_2\left(\frac{1}{5}\right) - \frac{4}{5}\log_2\left(\frac{4}{5}\right) = 0.722$

$\text{Entropy}(a1) = \frac{4}{9}(0.811) + \frac{5}{9}(0.722) = 0.763$

$\text{Gain}(a1) = 0.991 - 0.763 = 0.228$

a2:

$T \rightarrow P(+) = \frac{2}{4}, \; P(-) = \frac{2}{4}$

$F \rightarrow P(+) = \frac{2}{5}, \; P(-) = \frac{3}{5}$

$\text{Entropy}(T) = -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$

$\text{Entropy}(F) = -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.971$

$\text{Entropy}(a2) = \frac{4}{9}(1) + \frac{5}{9}(0.971) = 0.984$

$\text{Gain}(a2) = 0.991 - 0.984 = 0.007$

c) For $a3$, which is a continuous attribute, compute the information gain for every possible split.

2C) $a_3 = 1, 3, 4, 5, 6, 7, 8$

$4 \quad 7 \quad 9 \quad 11 \quad 13 \quad 15$

$t = 2, 3.5, 4.5, 5.5, 6.5, 7.5$

$t=2$:

$H(L) = 0$

$H(R) = -\frac{3}{8}\log_2\left(\frac{3}{8}\right) - \frac{5}{8}\log_2\left(\frac{5}{8}\right) = 0.954$

$E(2) = \frac{8}{9}(0.954) = 0.848$

$Gain(2) = 0.991 - 0.848 = \boxed{0.143}$

$t=3.5$:

$H(L) = -\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1$

$H(R) = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right) = 0.985$

$E(3.5) = \frac{2}{9}(1) + \frac{7}{9}(0.985) = 0.988$

$Gain(3.5) = 0.991 - 0.988 = \boxed{0.003}$

$t=4.5$:

$H(L) = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.918$

$H(R) = -\frac{2}{6}\log_2\left(\frac{2}{3}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right) = 0.918$

$E(4.5) = \frac{3}{9}(0.918) + \frac{6}{9}(0.918) = 0.918$

$Gain(4.5) = 0.991 - 0.918 = \boxed{0.073}$

$t=5.5$

$H(L)=-\frac{2}{5}\log_2(\frac{2}{5})-\frac{3}{5}\log_2(\frac{3}{5})=0.971$

$H(R)=-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})=1$

$E(5.5)=\frac{5}{9}(0.971)+\frac{4}{9}(1)=0.984$

$Gain(5.5)=0.991-0.984=\boxed{0.007}$

$t=6.5$

$H(L)=-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})=1$

$H(R)=-\frac{1}{3}\log_2(\frac{1}{3})-\frac{2}{3}\log_2(\frac{2}{3})=0.667$

$E(6.5)=\frac{6}{9}(1)+\frac{3}{9}(0.667)=0.973$

$Gain(6.5)=0.991-0.973=\boxed{0.018}$

$t=7.5$

$H(L)=-\frac{1}{2}\log_2(\frac{1}{2})-\frac{1}{2}\log_2(\frac{1}{2})=1$

$H(R)=0$

$E(7.5)=\frac{8}{9}(1)+\frac{1}{9}(0)=0.889$

$Gain(7.5)=0.991-0.889=\boxed{0.102}$

d) What is the best split (among a1, a2, and a3) according to the information gain?

a1 = 0.228
a2 = 0.007
a3(2.0) = 0.143

a1 has the best split because it has the largest information gain, meaning that splitting the data reduces the entropy or uncertainty. In other words, a1 provides the most information, and separates the good and bad examples.

e) What is the best split (between a1 and a2) according to the classification error rate?

2e) Classification error

a1:

$T = 1 - \frac{3}{4} = 0.25$

$F = 1 - \frac{4}{5} = 0.20$

$CE(a1) = \frac{4}{9}(0.25) + \frac{5}{9}(0.20) = 0.222$

a2:

$T = 1 - \frac{3}{5} = 0.4$

$F = 1 - \frac{2}{5} = 0.5$

$CE(a2) = \frac{4}{9}(0.5) + \frac{5}{9}(0.4) = 0.444$

a1 has the smaller value, meaning it has the better split because it produces fewer classifications overall, thus meaning that it is the better attribute by classification error rate.

f) What is the best split (between a1 and a2) according to the Gini index?

2-f) Gini Index

a1:

$T = 1-((\frac{3}{4})^2+(\frac{1}{4})^2) = 0.375$

$F = 1-((\frac{1}{5})^2+(\frac{4}{5})^2) = 0.32$

$G(a1) = \frac{4}{9}(0.375) + \frac{5}{9}(0.32) = 0.344$

a2:

$T = 1-(0.5^2+0.5^2) = 0.5$

$F = 1-(0.4^2+0.6^2) = 0.48$

$G(a2) = \frac{4}{9}(0.5) + \frac{5}{9}(0.48) = 0.488$

a1 has the better split for the Gini index as well because it has the smaller value out of the two Gini indexes, meaning that it will or does produce more pure splits.

g) Suppose you had access to 10× more training data. Do you think the best attribute split you chose in question f) would stay the same? Why or why not?

If we had access to 10 times more training data, then there is a chance that the best attribute split chosen in the previous question changes, but it is not guaranteed. This is because larger dataset would still reflect the same conditional distributions, meaning that the results would most likely remain the same.