Établissement : Institut Supérieure d'Informatique et de Technique de Communication

TD:

Decision Tree

Travail réalisé par:

Elagas Amel

3DNI G2

Année universitaire : 2021/2022

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

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

A. Compute the Gini index for the overall collection of training examples.

Answer:

This results in a single partition with 20 records and two possible classes with relative frequencies p and (1 − p), respectively. In this case, C0 and C1 have the same relative frequencies (p = 1 − p = 1/ 2)

Gini = 1 – p ^2 – (1 – p) ¨^2 = 2^p(1 – p) = 2p^ 2 = 2 ·*1/ 4 = 1 /2 = 0.5

B. Compute the Gini index for the ‘Customer ID‘attribute. Split the entire collection into 20 partitions based on the ‘Customer ID‘ attribute.

Answer:

Since each partition only contains a single record, it’s Gini index is zero by default. The weighted average of the Gini indices for all the partitions becomes:

Gini=0

C. Compute the Gini index for the Gender attribute.

Answer:

Split the entire collection into two partitions based on the Gender attribute (M or F). Each partition has 10 records. Set p as relative frequency of C0 in each case.

Gini(M) = 2*p*(1 − p) = 2 * 6 /10 ·*4 /10 = 48 /100 = 0.48

Gini(F) = 2*p*(1 − p) = 2 * 4 /10 *6 / 10 = 48 /100 = 0.48

Take the weighted averages of both Gini indices to determine the total Gini index for the given split.

Gini = 10/ 20 Gini(M) + 10 /20 Gini(F) = 48 /100 = 0.48

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

Answer:

This gives us 3 partitions (Family (4 records), Sports (8 records) and Luxury (8 records)).

Gini(Family) = 2 * 1 /4 * 3/ 4 = 0.375

Gini(Sports) = 2 * 8 / 8 * 0/ 8 = 0

Gini(Luxury) = 2 * 1/ 8 *7 8 = 0.2656

So, the weighted average of these indices is:

Gini = 4 / 20 * Gini(Family) + 8 /20 * Gini(Sports) + 8 /20 * Gini(Luxury)
= 4 / 20 * 6 / 16 + 8 /20 *14/ 64
= 0.16

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

Answer:

Here we get 4 partitions (Small (5 records), Medium (7 records), Large (4 records) and Extra Large (4 records)).

Gini(Small) = 2 * 3 /5 *2 /5 = 0.48

Gini(Medium) = 2 * 3 /7 * 4 /7 = 0.4898

Gini(Large) = 2 * 2 / 4 * 2 / 4 = 0.5

The weighted averages of these indices is:

Gini = (4 /20 * 12/ 25 + 8/ 20 * 24 /49 + 4 /20 * 1/ 2 + 4 /20 * 1 /2) = 0.49

F. Which attribute is better, Gender, Car Type, or Shirt Size?

Answer:

Let us determine the gain from each split:

Gain(Gender) = 0.5 − 0.48 = 0.02
Gain(Car Type) = 0.5 − 0.1625 = 0.3375
Gain(Shirt Size) = 0.5 − 0.4919 = 0.0081

From the results above, the Car Type would give the highest gain in purity (based on the Gini index).

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

Answer:

By definition, the Gini index of each ID-partition is zero, since it only contains a single record.
Adding more IDs to the table will only increase the number of partitions, resulting in no further purity gain.

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 class attribute?

Answer:

We have a single partition with 9 records.

Let p be the relative frequency of + (hence, 1 − p for −).

Entropy (Class) = p log2 (p) − (1 − p) log2 (1 − p)

So,

Entropy (Class) = p log2 (p) − (1 − p) log2 (1 − p)

= − 4 /9 log2 (4/ 9) – 5/ 9 log2 (5 /9)

= 0.9911

B. What are the information gains of A1 and A2 relative to these training examples?

Answer:

The set is split into 2 partitions in both cases. Start with a1. Putting all instances into a table:

A_1	**+**	**-**
T	**3**	**1**
F	**1**	**4**

➔

A1	+	-
T	3/4	1/4
F	1/5	4/5

Entropy (A1) = 4/9 *(-3/4*log(3/4)- ¼*log(1/4)) + 5/9 * (-1/5* log(1/5) – 4/5 *log (4/5))
=0.76

Similarily for A2:

A2	+	-
T	2	3
F	2	2

➔

A2	+	-
T	2/5	3/5
F	2/4	2/4

Entropy (A2) = 5/9 *(-2/5*log(2/5)- 3/5 *log(3/5)) + 4/9 * (-2/4* log(2/4) – 2/4 *log (2/4))
=0.98

Gain (A1) = 0.9911 − 0.7616 = 0.229
Gain (A2) = 0.9911 − 0.9838 = 0.007

C. For A3, which is a continuous attribute, compute the information gain for every possible split.

Answer:

Value of A3 that occur in the given table are in the following range of [1.0, 8.0]. After sorting, we'll set split positions midway between neighboring values. Table sorted by A3 (then by ID):

Instance	A3	Target Class
1	1.0	+
6	3.0	-
4	4.0	+
3	5.0	-
9	5.0	-
2	6.0	+
5	7.0	-
8	7.0	+
7	8.0	-

Below, split positions of a3 are displayed in the top left corner of each count matrices.

Below each matrix is the weighted entropy and information gain (E/G) in each case.

The split position with maximal information gain is emphasised in bold.

This corresponds to splitting at A3 = 2.0

0.5	<=	>
+	0	4
-	0	5
E/G	0.9911	0

2.0	<=	>
+	1	3
-	0	5
E/G	**0.8484**	**0.1427**

3.5	<=	>
+	1	3
-	1	4
E/G	0.9858	0.0026

4.5	<=	>
+	2	2
-	1	4
E/G	0.9183	0.0728

5.5	<=	>
+	2	2
-	3	2
E/G	0.9839	0.0072

6.5	<=	>
+	3	1
-	3	2
E/G	0.9728	0.0183

7.5	<=	>
+	4	0
-	4	1
E/G	0.8889	0.1022

8.5	<=	>
+	4	0
-	5	0
E/G	0.9911	0

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

Answer:

By maximizing information gain:

Gain (a1) = 0.2294
Gain (a2) = 0.0072
Gain (a3) = 0.1427

we get the best split from a1.

E. What is the best split (between A1 and A2) according to the misclassification error rate?

Answer:

We've already calculated the relative frequencies: Start with A1:

P(A1)	+	-
T	3/4	1/4
F	1/5	4/5

➔ **Error(A1) = 4 /9 (1 – 3 /4) + 5/ 9 (1 – 4 /5) = 2/ 9**

P(A2)	+	-
T	2/5	3/5
F	2/4	2/4

➔ **Error(A2) = 5 /9 (1 – 3/ 5) + 4/ 9 (1 – 2 /4) = 4/ 9**

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

Answer:

P(A1)	+	-
T	3/4	1/4
F	1/5	4/5

➔ **Gini(A1) = 4 /9 (1 – (3^2 /4^2) - (1^2/4^2)) + 5/ 9 (1 – (1^2 /5^2) - (4^2/5^2)) = 0.34**

P(A2)	+	-
T	2/5	3/5
F	2/4	2/4

➔ **Gini(A2) = 5/9 (1 – (2^2 /5^2) - (3^2/5^2)) + 4/ 9 (1 – (2^2 /4^2) - (2^2/4^2)) = 0.48**