

Classification Discussion question 1 – Gini Index:

Consider the training examples shown in Table 1 for a binary classification problem.

Table 1. Data set

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

Questions:

- Compute the Gini index for the overall collection of training examples.
- Compute the Gini index for the Customer ID attribute.
- Compute the Gini index for the Gender attribute.
- Compute the Gini index for the Car Type attribute using multiway split.
- Compute the Gini index for the Shirt Size attribute using multiway split.
- Which attribute is better, Gender, Car Type, or Shirt Size?
- Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.