Nitte Meenakshi Institute of Technology
Department of Computer Science & Engineering

18CS54 – Data Mining

1. Consider the training examples shown in **Table 1** for a binary classification problem.
(a) Compute the Gini index for the overall collection of training examples.
(b) Compute the Gini index for the Customer ID attribute.
(c) Compute the Gini index for the Gender attribute.
(d) Compute the Gini index for the Car Type attribute using multiway split.
(e) Compute the Gini index for the Shirt Size attribute using multiway split.
(f) Which attribute is better, Gender, Car Type, or Shirt Size?
(g) Explain why Customer ID should not be used as the attribute test condition even though it has the lowest Gini.

Table 1. Sample Dataset1

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

$$GINI(t) = 1 - \sum_{j} [p(j \mid t)]^2$$

$$GINI_{split} = \sum_{i=1}^{k} \frac{n_i}{n} GINI(i)$$

Answers:

a. 0.5
b. The gini for each Customer ID value is 0. Therefore, the overall gini for Customer ID is 0
c. The gini for Male is = 0.5. The gini for Female is also 0.5.
   Therefore, the overall gini for Gender is 0.5 x 0.5 + 0.5 x 0.5 = 0.5.
d. The gini for Family car is 0.375, Sports car is 0, and Luxury car is
   0.2188. The overall gini is 0.1625.
e. The gini for Small shirt size is 0.48, Medium shirt size is 0.4898,
   Large shirt size is 0.5, and Extra Large shirt size is 0.5. The overall gini for

Shirt Size attribute is 0.4914.

f. Car Type because it has the lowest gini among the three attributes.

g. The attribute has no predictive power since new customers are assigned to new Customer IDs.

2. Consider the training examples shown in **Table 2** for a binary classification problem.

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

(b) What are the information gains of a_1 and a_2 relative to these training examples?

(c) For a_3, which is a continuous attribute, compute the information gain for every possible split.

(d) What is the best split (among a_1, a_2, and a_3) according to the information gain?

(e) What is the best split (between a_1 and a_2) according to the classification error rate?

(f) What is the best split (between a_1 and a_2) according to the Gini index?

Table 2. Sample Dataset2

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	−

Source: Pang-Ning Tan, Vipin Kumar, Michael Steinbach: Chapter 4-Exercises, **Introduction to Data Mining**, Pearson, 2012.