

# Data Mining

## Week 7 Exercise Sheet

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

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	Small	C1

- (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 **Short Size**?
- (g) Explain why **Customer ID** should not be used as the attribute test condition even though it has the lowest Gini.

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

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	+
0	F	T	5.0	-

- (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, compute the information gain for every possible split.
- (d) What is the best split (among  $a_1, a_2, a_3$ ) according to the information gain?
- (e) What is the best split (between  $a_1, a_2$ ) according to the classification error rate?
- (f) What is the best split (between  $a_1, a_2$ ) according to the Gini index?