HW2
Deadline: 11/6

1. 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

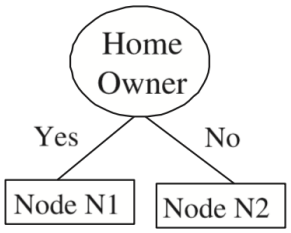
a) (5%) Compute the entropy for the overall collection of training examples.
b) (15%) What is the information gain when splitting on the **Gender** attribute?
c) (20%) What is the information gain when splitting on the **Car Type** attribute (using multiway split)?
d) (20%) What is the information gain when splitting on the **Shirt Size** attribute (using multiway split)?
e) (3%) According to the information gain, which attribute is better, **Gender**, **Car Type**, or **Shirt Size**?

2. For a binary classification problem, the composition of labeled training instances at a parent node is summarized in the Table below.

	Parent
No	**7**
Yes	**3**

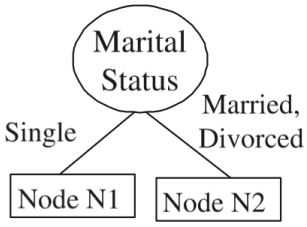
a) (5%) What is the Gini index of the parent node?

b) (16%) Consider splitting a parent node into two child nodes, N1 and N2, using the **Home Owner** attribute. The composition of labeled training instances at each child node is summarized in Figure 1. What is the Gini index of each child node? What is the gain in the Gini index when splitting on this attribute? According to this impurity measure, will you consider this attribute test condition?



	N1	**N2**
No	**3**	**4**
Yes	**0**	**3**

Figure 1.



	N1	**N2**
No	**3**	**4**
Yes	**2**	**1**

Figure 2.

c) (16%) Consider splitting a parent node into two child nodes, N1 and N2, using the **Marital Status** attribute. The composition of labeled training instances at each child node is summarized in Figure 2. What is the Gini index of each child node? What is the gain in the Gini index when splitting on this attribute? According to this impurity measure, will you consider this attribute test condition?