Data Classification (1)

1. 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.
b) Compute the Gini index for the **Gender** attribute.
c) Compute the Gini index for the **Car Type** attribute using multiway split.
d) Compute the Gini index for the **Shirt Size** attribute using multiway split.
e) Which attribute is better, **Gender**, **Car Type**, or **Shirt Size**?

2. Consider the one-dimensional data set shown in the following table:

x	0.5	0.3	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	-	-	+	+	+	-	-	+	-	-

a) Classify the data point x=5.0 according to its 1-, 3-, 5-, and 9- nearest neighbors (using majority vote).
b) Repeat the previous analysis using the distance-weighted voting approach.

3. Choose **Iris data set** (you can use your course project data set) and use R to generate decision tree out of it. You need to submit the screenshot showing the decision trees and explain the classification results, or use R studio to compile report.