Homework 4

1.Exercises

1.1 Chapter 3

Question 2:

a). Gini Index = $1-2*0.5^2 = 0.5$
b). Gini Index for Customer ID is 0.
c). GM = GI of Gender Male = 0.48
GF = GI of Gender Female = 0.48
GI of both Genders = GM * total fraction of male + GF * total fraction of female
= 0.48
d). GI of Car Type = GI of family car* total fraction of family car + GI of Sport car * total fraction of Sport + GI of Luxury car * total fraction of Luxury car
= 0.375 *0.2 + 0* 0.45 + 0.2188*0.4
= 0.16252
e). GI of Small size = 0.48
GI of Medium size = 0.48
GI of Large size = 0.5
GI of Extra Large size = 0.5
Total GI of Size = 0.491
f). As Car Type has lowest GI value It is a better attribute
g). Customer ID is used to uniquely identify a customer. It doesn't have any value to predict.

Question 3:

a). Entropy of training examples = $-4/9 \log 2 (4/9) - 5/9 \log 2 (5/9) = 0.9911$
b).
Entropy for a1 =
$4/5 (-(3/4)\log_2(3/4)-(1/4)\log_2(1/4)) +5/9 (-(1/5)\log_2(1/5) – (4/5)\log_2(4/5))$
=0.716
IG of a1 = 0.229

Entropy for a2 =
$5/9 (-(2/5)\log_2(2/5)-(3/5)\log_2(3/5)) +4/9 (-(2/4)\log_2(2/4) – (2/4)\log_2(2/4))$
=0.983.
IG of a2 = 0.007

c). After reordering the table based on a3 values and obtaining split values. As the IG is highest at split point 2. Best split is at 2.
d). When we compare IG values of a1, a2, a3. a1 has Highest IG. So, a1 gives best split.

e). for a1 error rate is = 0.222
for a2 error rate is = 0.444
a1 gives best split
f). GI of a1 = 0.344
GI of a2 = 0.488
a1 gives the best split.

Question 5:

a). The IG on split A is 0.281
The IG on Split B is 0.256
So, A is selected for best split.

b). IG on Gini on split A is 0.137
IG on Gini on split b is 0.163
Based on IG on Gini Index, B is selected for best split.

c). Yes, even though these measures have similar range and monotonous behavior, their respective gains, they do not necessarily behave in the same way.

1.2 Chapter 4

Question 18:

a). (P(error) = P(error|+) * P(+) + P(error|-) * P(-) = 0.50 * 0.50 + 0.50 * 0.50 = 0.50)
The expected error rate of the classifier on the test data is 50%
b). (P(error) = P(error | +) P(+) + P(error| -) P(-) = 0.2 * 0.5 + 0.8 * 0.5 = 0.1 + 0.4 = 0.5)
The expected error rate of the classifier on the test data with given probabilities is 50%
c). The expected error of a classifier that predicts every test record to be positive is 33%
d). The expected error of a classifier that predicts every test record to be positive with given probabilities is 44.4%