CS 422 Homework4

Xingli Li, Illinois Institute of Technology

1. Exercises

Problem 1.1-(2)

a. Compute the Gini index for the overall collection of training examples.

```
Gini=0.5
```

b. Compute the Gini index for the **Customer ID** attribute

```
Gini=0
```

c. Compute the Gini index for the **Gender** attribute.

```
Gini(Female)=0.5
Gini(Male)=0.5
Gini=0.5
```

d. Compute the Gini index for the **Car Type** attribute using multiway split.

```
Gini(Family)=0.375
Gini(Sport)=0
Gini(Luxury)=0.219
Gini = 0.163
```

e. Compute the Gini index for the **Shirt Size** attribute using multiway split.

```
 Gini(Small)= 0.48
 Gini(Medium)=0.490
 Gini(Large)=0.5
 Gini(Extra Large)=0.5
 Gini=0.491
```

f. Which attribute is better, **Gender**,**Car Type** , or **Shirt Size** ?

```
Car Type is better.
```

g. Explain why **Customer ID** should not be used as the attribute test condition even though it has the lowest Gini.

```
Having a low impurity value alone is insufficient to find a good attribute test condition for a node,
if an attribute produces a large number of splits, its split information is also large, which in turn,
reduces the gain ratio.
```

Problem 1.1-(3)

a. What is the entropy of this collection of training examples with respect to the class attribute?

```
Entropy = 0.991
```

b. What are the information gains of a1 and a2 relative to these training examples?

```
Entropy(a1) = 0.762
IGain(a,a1) = 0.991-0.762= 0.229
IGain(a,a2) = 0.991-0.984= 0.007
```

c. For a3 , which is a continuous attribute, compute the information gain for every possible split.

a3	1.0	3.0	4.0	5.0	5.0	6.0	7.0	7.0
class label	+	-	+	-	-	+	+	-
split point	2.0	3.5	4.5	5.5	5.5	6.5	7.5	7.5
entropy	0.848	0.989	0.918	0.984	0.984	0.973	0.889	0.889
info gain	0.143	0.002	0.073	0.007	0.007	0.018	0.102	0.102

```
From above table, I know that the best split is at split point equals to 2.
```

d. What is the best split (among a1 , a2 and a3) according to the information gain?

```
a1
```

e. What is the best split (between a1 and a2) according to the misclassification error rate?

```
a1
```

f. What is the best split (between a1 and a2) according to the Gini index?

```
Gini(a1)=0.344
Gini(a2)=0.489
The best split is a1.
```

Problem 1.1-(5)

a. Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

```
Entropy = 0.971

IGain(A)= 0.281
IGain(B)= 0.257

Because IGain(A) > IGain(B), I will choose A attribute.
```

b. Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

```
Gini = 0.48

GGain(A)= 0.137
GGain(B)= 0.163

Because GGain(B) > GGain(A), I will choose B attribute.
```

c. Figure 3.11 shows that entropy and the Gini index are both monotonically increasing on the range [0, 0.5] and they are both monotonically decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes? Explain.

```
Yes, even though these measures have similar range and monotonous,
they are scaled differences of the measures, Gini index is based on quadratic
equation,
information gain is based on Log function.
```

Problem 1.2-(18)

(show your work)

Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains instances from two classes, “ ” and “ .” Half of the data set is used for training while the remaining half is used for testing.

a. Suppose there are an equal number of positive and negative instances in the data and the decision tree classifier predicts every test instance to be positive. What is the expected error rate of the classifier on the test data?

Assume let n is total test date set size, acoording probability of binomial distribution,

$(n * \frac{1}{2} + 0 * \frac{1}{2}) * \frac{1}{n} = 0.5$

```
Half of the test data here are classified incorrectly.
Error rate = 0.5.
```

b. Repeat the previous analysis assuming that the classifier predicts each test instance to be positive class with probability 0.8 and negative class with probability 0.2.

$(0.2n * \frac{1}{2} + 0.8n * \frac{1}{2}) * \frac{1}{n} = 0.5$

```
Error rate =0.5
```

c. Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. What is the expected error of a classifier that predicts every test instance to be positive?

$(n * \frac{2}{3} + 0 * \frac{1}{3}) * \frac{1}{n} = 0.667$

```
Error rate = 0.667
```

d. Repeat the previous analysis assuming that the classifier predicts each test instance to be positive class with probability 2/3 and negative class with probability 1/3.

$$(\frac{2n}{3} * \frac{2}{3} + \frac{n}{3} * \frac{1}{3}) * \frac{1}{n} = 0.556$$

```
Error rate = 0.556
```

Problem 1.3

Multiclass classification

Using the confusion matrix from multiclass.Rmd notebook (from Lecture 7), create a binary-class confusion matrix using the “one-vs-many” strategy for each class. Then, for each class, compute the sensitivity, specificity and precision to two decimal places. Show all work, including the binary class confusion matrices.

Confusion Matrix and Statistics: （row: prediction; column: actual）

	setosa (-)	**versicolor, virginica (+)**
setosa (-)	10 (TN)	0 (FN)
versicolor,virginica (+)	0 (FP)	20 (TP)

```
sensitivity = TPR = TP/(TP+FN) = 1.0
specificity = TNR = RN/(TN+FP) = 1.0
precision = TP/(TP+FP) = 1.0
```

	setosa，versicolor(+)	**virginica(-)**
setosa，versicolor(+)	20 (TP)	1 (FP)
virginica(-)	0 (FN)	9 (TN)

```
sensitivity = 1.00
specificity = 0.90
precision =  20/21 = 0.95
```

	versicolor(-)	**virginica，setosa(+)**
versicolor(-)	10 (TN)	0 (FN)
virginica，setosa(+)	1 (FP)	19 (TP)

```
sensitivity = 0.95
specificity = 1.00
precision = 19/20 = 0.95
```