Name : Pavani Rangineni

CWID: A20516359

DATA MINING HW4

1.1.2.Consider the training examples shown in Table for a binary classification problem.

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

Gini = 1 − 2 × 0.5 2 = 0.5.

(b) Compute the Gini index for the Customer ID attribute.

The Gini for each Customer ID value is 0. Therefore, the overall gini for Customer ID is 0.

(c) Compute the Gini index for the Gender attribute.

Gini index for male = $1-(6/10)^2-(4/10)^2$ = 1-36/100-16/100 = 0.48

Gini index for female = $1-(4/10)^2 -(6/10)^2$ = 1-16/100-36/100 = 0.48

Overall gini = 0.48*0.5 + 0.48*0.5 = 0.48

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

The Gini for Family car is 0.375, Sports car is 0, and Luxury car is 0.2188.

The overall Gini is 0.1625.

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

The gini for Small shirt size is 0.48, Medium shirt size is 0.4898, Large shirt size is 0.5, and Extra Large shirt size is 0.5.

The overall gini for Shirt Size attribute is 0.4914.

(f) Which attribute is better, Gender, Car Type, or Shirt Size?

Car Type because it has the lowest gini among the three attributes.

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

The attribute has no predictive power since new customers are assigned to new Customer IDs.

1.1.3.Consider the training examples shown in Table 4.2 for a binary classification problem.

(a) What is the entropy of this collection of training examples with respect to the positive class?

There are four positive examples and five negative examples. Thus, P (+) = 4/9 and P (−) = 5/9.

The entropy of the training examples is −4/9 log 2 (4/9) − 5/9 log 2 (5/9) = 0.9911.

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

The entropy of a1 = 4/9[-(3/4) log2(3/4)-(1/4) log2(1/4)] + 5/9[-(1/5)log2(1/5)-(4/5)log2(4/5)] = 0.7616

The information gain of a1 = 0.9911 – 0.7616 = 0.2294

The entropy of a2 = 5/9[-(2/5) log2(2/5)-(3/5) log2(3/5)] + 4/9[-(2/4)log2(2/4)-(2/4)log2(2/4)] = 0.9839

The information gain of a2 = 0.9911 – 0.9839 = 0.0072

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

The best split for a 3 occurs at split point equals to 2

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

According to information gain, a 1 produces the best split.

(e) What is the best split (between a 1 and a 2) according to the classification error rate?

For attribute a 1 : error rate = 2/9. For attribute a 2 : error rate = 4/9. Therefore, according to error rate, a 1 produces the best split.

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

The gini index of a1 = $4/9[1-(3/4)^2-(1/4)^2] + 5/9[1-(1/5)^2-(4/5)^2] = 0.3444$

The gini index of a2 = $5/9[1-(2/5)^2-(3/5)^2] + 4/9[1-(2/4)^2-(2/4)^2] = 0.4880$

Since the gini index of a1 is smaller. So a1 has the better split.

1.1.5. Consider the following data set for a binary class problem.

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

The overall entropy before splitting is $E_{orig\,=}$ -0.4log0.4 – 0.6 log0.6 = 0.9710

The information gain after splitting on A Δ = E_{orig}- 7/10$E_{A=T}$ – 3/10$E_{A=F}$= 0.2813

The information gain after splitting on A Δ = E_{orig}- 4/10$E_{B=T}$ – 6/10$E_{B=F}$= 0.2565

Therefore A will be chosen.

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

The overall gini before splitting is $G_{orig} = 1-0.4^2-0.6^2=0.48$

The gain in gini after splitting on A Δ = G_{orig}- 7/10$G_{A=T}$ – 3/10$G_{A=F}$= 0.1371

The gain in gini after splitting on B Δ = G_{orig}- 4/10$G_{A=T}$ – 6/10$G_{A=F}$= 0.1633

Therefore, attribute B will be chosen.

(c) Figure 4.13 shows that entropy and the Gini index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini index favor different attributes?

Yes, even though these measures have similar range and monotonous behavior, their respective gains, Δ, which are scaled differences of the measures, do not necessarily behave in the same way.

1.2.18. 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 records 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 records in the data and the decision tree classifier predicts every test record to be positive. What is the expected error rate of the classifier on the test data?

50%. (P(error) = P(error|+) * P(+) + P(error|-) * P(-) = 0.50 * 0.50 + 0.50 * 0.50 = 0.50)

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

50%. (P(error) = P(error | +) P(+) + P(error| -) P(-) = 0.2 * 0.5 + 0.8 * 0.5 = 0.1 + 0.4 = 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 record to be positive?

33%.

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

44.4%. (P(error) = P(error|+) * P(+) + P(error|-) * P(-) = (0.33 *0.67)+(0.67*0.33) = 0.4422)

1.3 Multi-class problem (The numbers below may be different from the lecture, but that should not matter for the purpose of understanding the process.)

		Actual		
		Setosa	Versicolor	Virginica
Predicted	Setosa	8	0	0
	Versicolor	0	10	1
	Virginica	0	2	9

For class Setosa:

		Actual	
		Setosa	{Versicolor,Virginica}
Predicted	Setosa	8	0
	{Versicolor,Virginica}	0	22

Note that for the TNs is the sum of instances where = actual is Versicolor and predicted is Versicolor + actual is Versicolor and predicted is Virginica + actual is Virginica and predicted is Virginica +actual is Virginica and predicted is Versicolor = 10 + 1 + 2 + 9 = 22

Sensitivity = 8/(8+0) = 8/8 = 1.0
Specificity = 22/(22+0) = 1.0
Precision = 8/(8+0) = 1.0
For class Versicolor:

Predicted	Actual		
		Versicolor	{Setosa,Virginica}
	Versicolor	10	1
	{Setosa,Virginica}	2	8+0+0+9=17

Sensitivity = 10/(10+2) = 0.83
Specificity = 17/(17+1) = 0.94
Precision: 10/(10+1) = 0.91
For class Virginica:

Predicted	Actual		
		Virginica	{Versicolor,Setosa}
	Virginica	9	2+0 = 2

Sensitivity = 9/(9+1) = 0.90
Specificity = 18/(18+2) = 0.90
Precision = 9/(9+2) = 0.82