**Homework 3: Classification**

Q1.) Node Impurity Questions.

1. What is the entropy of the collection of the training examples?

Ans.) In the given data, there are four ‘+’ class labels and five ‘-‘ class labels. i.e., p(+) = 4/9 and p(-) = 5/9.

We know that **Entropy(S)** =

Therefore, the entropy of the training examples is, -4/9 log2 (4/9) + (-5/9) log2 (5/9) which is

-4/9 log2 (4/9) – 5/9 log2 (5/9) = 0.9911.

b.) What is the information gain of a1 relative to the training examples?

Ans.) We know that **Information Gain** = **Entropy before splitting - Entropy after splitting.**

We also know that Entropy before splitting is nothing but the entropy value of the collection of the training example. And the entropy after splitting means the value of the entropy of the a1 attribute.

Entropy Before Splitting: 0.9911 (we already did it in the above problem).

Entropy After Splitting: we need to calculate the entropy of the attribute a1.

For attribute a1, the corresponding counts and probabilities are:

|  |  |  |
| --- | --- | --- |
| a1 | + | - |
| T | 3 | 1 |
| F | 1 | 4 |

The entropy for a1 is, +

= 0.7616.

Therefore, the information gain for a1 is 0.9911- 0.7616 = 0.2294.

1. What is the best split (between a1 and a3) according to the information gain?

Ans.) We know that **Information Gain** = **Entropy before splitting - Entropy after splitting.**

We also know that Entropy before splitting is nothing but the entropy value of the collection of the training example. And the entropy after splitting means the value of the entropy of the a3 attribute.

Entropy Before Splitting: 0.9911 (From the problem (a)).

Entropy After Splitting: we need to calculate the entropy of the attribute a3.

For attribute a3, the corresponding counts and probabilities are:

|  |  |  |
| --- | --- | --- |
| a3 | + | - |
| <= 3,5 | **1** | **1** |
| > 3,5 | **3** | **4** |

The entropy for a3 is, + = 0.766.

Therefore the information gain for a3 is 0.9911 – 0.766 = 0.2251.

Therefor the best split according to the information gain is a1.

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

Ans.) We know that **Gini(P)** =

**Gini(P)** = **1 –**

The Gini index for attribute a1 is:

The Gini index for the attribute a2 is:

Since the Gini Index for the attribute a1 is smaller than that of the attribute a2, hence it produces the better split.

Q2.) Decision Tree Construction.

1. According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

Ans.) We know that for Classification Error, **Error(t)** = **1 – maxi p (i / t)**

Now we calculate the error rate for the training dataset before splitting the attributes,

E = 1 – max .

Now we calculate the error rate for the training dataset after splitting the attributes,

Contingency table of Attribute A:

|  |  |  |
| --- | --- | --- |
| Class Label(Instances) | A=T | A=F |
| + | 25 | 25 |
| - | 0 | 50 |

After splitting on attribute A, the gain in error rate is:

EA=T = 1 – max(25/25 , 0/25) = 0/25 = 0

EA=F = 1 – max(25/75 , 50/75) = 25/75

As we know, Gain = E(before splitting) – E(after splitting),

ΔA = E – 25/100 EA=T – 75/100 EA=F = 25/100.

Contingency table of Attribute B:

|  |  |  |
| --- | --- | --- |
| Class Label(Instances) | B=T | B=F |
| + | 30 | 20 |
| - | 20 | 30 |

After splitting on attribute B, the gain in error rate is:

EB=T = 1 – max(30/50 , 20/50) = 20/50

EB=F = 1 – max(20/50 , 30/50) = 20/50

As we know, Gain = E(before splitting) – E(after splitting),

ΔB = E – 50/100 EB=T – 50/100 EB=F = 10/100.

Contingency table for Attribute C:

|  |  |  |
| --- | --- | --- |
| Class Label(Instances) | C=T | C=F |
| + | 25 | 25 |
| - | 25 | 25 |

After splitting on attribute C, the gain in error rate is:

EC=T = 1 – max(25/50 , 25/50) = 25/50

EC=F = 1 – max(25/50 , 25/50) = 25/50

As we know, Gain = E(before splitting) – E(after splitting),

ΔC = E – 50/100 EC=T – 50/100 EC=F = 0/100 = 0.

Therefore according to the classification error rate, the attribute A would be chosen as the first splitting attribute because it has the highest gain.

1. Split the two children of the root node.

Ans.) For the attribute A=T child node, is pure from the above problem. No further splitting is required, as for A=F child node, the classification error of the A=F child node is (from the above problem):

EA=F = 25/75 (from problem(a)) 🡪 let it be E = 25/75.

Now the distribution of the training instances is:

|  |  |  |  |
| --- | --- | --- | --- |
| B | C | + | - |
| T | T | 0 | 20 |
| F | T | 0 | 5 |
| T | F | 25 | 0 |
| F | F | 0 | 25 |

Contingency table for Attribute B is:

|  |  |  |
| --- | --- | --- |
| Class Label(Instances) | B=T | B=F |
| + | 25 | 30 |
| - | 20 | 0 |

After splitting on attribute B, the gain in error rate is:

EB=T = 1 – max(25/45 , 20/45) = 20/45

EB=F = 1 – max(30/30 , 0/30) = 0

As we know, Gain = E(before splitting) – E(after splitting),

ΔB = E – 45/75 EB=T – 20/75 EB=F = 5/75.

Contingency table for Attribute C is:

|  |  |  |
| --- | --- | --- |
| Class Label(Instances) | C=T | C=F |
| + | 0 | 25 |
| - | 25 | 25 |

After splitting on attribute C, the gain in error rate is:

EC=T = 1 – max(0/25 , 25/25) = 0/25

EC=F = 1 – max(25/50 , 25/50) = 25/50

As we know, Gain = E(before splitting) – E(after splitting),

ΔC = E – 25/75 EC=T – 50/75 EC=F = 0.

Therefore according to the classification error rate, the attribute B would be chosen as a splitting attribute because it has the highest gain.

Q3.) Classification Model Evaluation.

1. Plot the ROC curve for both M1 and M2 (You should plot them on the same graph). Which model do you think is better. Explain your reasons. Manually show the steps for calculating TP, FP, TN, FN, TPR, FPR, and draw ROC curve for M1.

Ans.) First, we have to sort for each instance P(+|A) of both the models M1 and M2 in ascending order(small to big)

**For M1:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| True Class | - | - | - | - | + | + | - | + | + | + |
| P(+|A) | **0.08** | **0.15** | **0.35** | **0.44** | **0.45** | **0.47** | **0.55** | **0.67** | **0.69** | **0.73** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TP | 5 | 5 | 5 | 5 | 5 | 4 | 3 | 3 | 2 | 1 |
| FP | **5** | **4** | **3** | **2** | **1** | **1** | **1** | **0** | **0** | **0** |
| TN | **0** | **1** | **2** | **3** | **4** | **4** | **4** | **5** | **5** | **5** |
| FN | **0** | **0** | **0** | **0** | **0** | **1** | **2** | **2** | **3** | **4** |
| TPR | **1** | **1** | **1** | **1** | **1** | **0.8** | **0.6** | **0.6** | **0.4** | **0.2** |
| FPR | **1** | **0.8** | **0.6** | **0.4** | **0.2** | **0.2** | **0.2** | **0** | **0** | **0** |

* Then apply threshold at each unique value of P(+|A), i.e., count the no. of TP, FP, TN, FN at each threshold δ.
* Assign the selected record with P >= δ to the positive class.
* Assign the selected record with P < δ to the negative class.
* Then calculate the True Positive Rate (TPR) and False Positive Rate (FPR).

**ROC Curve for M1:**

**For M2:**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| True Class | + | + | - | - | + | - | - | + | + | - |
| P(+|A) | **0.01** | **0.03** | **0.04** | **0.05** | **0.09** | **0.31** | **0.38** | **0.45** | **0.61** | **0.68** |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| TP | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 1 | 0 |
| FP | **5** | **5** | **5** | **4** | **3** | **3** | **2** | **1** | **1** | **1** |
| TN | **0** | **0** | **0** | **1** | **2** | **2** | **3** | **4** | **4** | **4** |
| FN | **0** | **1** | **2** | **2** | **2** | **3** | **3** | **3** | **4** | **5** |
| TPR | **1** | **0.8** | **0.6** | **0.6** | **0.6** | **0.4** | **0.4** | **0.4** | **0.2** | **0** |
| FPR | **1** | **1** | **1** | **0.8** | **0.6** | **0.6** | **0.4** | **0.2** | **0.2** | **0.2** |

**ROC Curve for M2:**

**ROC Curve for Both M1 and M2:**

The model M1 is better, because it has a higher Theoretical value than that of model M2.

The ROC Curves that we get from the manual calculation and the code output are different.

1. For model M1, suppose you choose the cut-off threshold to be t = 0.5. In other words, any test instances whose posterior probability is greater than ‘t’ will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value.

Ans.) Contingency table for M1 at threshold (δ) = 0.5 is:

|  |  |  |
| --- | --- | --- |
| True Class | P(+|A)  + | - |
| + | **TP = 3** | **FN = 2** |
| - | **FP = 1** | **TN = 4** |

Now we have to calculate the performance metrics,

We know that, for Precision =  **=** 3/3+1 = 0.75.

And as for Recall =  **=** 3/3+2 = 0.6.

And the F-measure = = 2/1.66+1.33 = 0.66.

1. Repeat part (b) for Model M1 using the threshold t = 0.1. Which threshold do you prefer, t = 0.5 or t = 0.1? Are the results consistent with what you expect from the ROC curve?

Ans.) Contingency table for M1 at threshold (δ) = 0.1 is:

|  |  |  |
| --- | --- | --- |
| True Class | P(+|A)  + | - |
| + | **TP = 5** | **FN = 0** |
| - | **FP = 4** | **TN = 1** |

Now we have to calculate the performance metrics,

We know that, for Precision =  **=** 5/5+4 = 0.55.

And as for Recall =  **=** 5/5+0 = 1.

And the F-measure = = 2/1+1.81 = 0.71.

When comparing the Threshold values of 0.1 and 0.5, threshold value 0.1 indicates a larger F- measure which concludes that it is the preferred threshold. And yes, the result is consistent with what was found in the part (a) of the ROC curve.