1. **Decision Trees**

* Total 3 attributes (Department, Age, Salary)
  + Department {Sales, Systems, Marketing}
  + Age {21-30, 31-40, 41-50}
  + Salary {Low, Medium, High}
* Class label variable: Status {Junior, Senior}
* Total records = 12
* Using Multiway split when constructing the decision tree

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Age** | **Salary** | **Status** |
| Sales | 31 – 40 | Medium | Senior |
| Sales | 21 – 30 | Low | Junior |
| Sales | 31 – 40 | Low | Junior |
| Systems | 21 – 30 | Medium | Junior |
| Systems | 31 – 40 | High | Senior |
| Systems | 21 – 30 | Medium | Junior |
| Systems | 41 – 50 | High | Senior |
| Marketing | 31 – 40 | Medium | Senior |
| Marketing | 31 – 40 | Medium | Junior |
| Marketing | 41 – 50 | High | Senior |
| Marketing | 21 – 30 | Low | Junior |
| Marketing | 21 – 30 | Medium | Junior |

Using data given in Table 1 as training data, answer the following question:

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Age** | **Salary** | **Status** |
| Sales | 21 – 30 | High |  |
| Systems | 21 – 30 | Medium |  |
| Marketing | 41 – 50 | High |  |
| Marketing | 31 – 40 | Low |  |

1. Construct decision tree using Gini index. Show all work and draw the resulting tree (no pruning)

**Ans:**

Gini index of **Department**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Sales** | **Systems** | **Marketing** |
| Junior | 1 | 2 | 2 |
| Senior | 2 | 2 | 3 |

Gini index of **Age**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Age** | **21 – 30** | **31 – 40** | **41 – 50** |
| Junior | 5 | 2 | 0 |
| Senior | 0 | 3 | 2 |

Gini index of **Salary**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Salary** | **Low** | **Medium** | **High** |
| Junior | 3 | 4 | 0 |
| Senior | 0 | 2 | 3 |

Summarizing,

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Gini** | **Gain** |
| Department | 0.4778 | 0.0084 |
| Age | 0.2 | **0.2861** |
| Salary | 0.2222 | 0.2639 |

Gain(Age) is greater among all the attributes. Thus, selected attribute for splitting is ‘Age’.

C:\Users\foo\AppData\Local\Microsoft\Windows\INetCacheContent.Word\Q1_2.png

Now, data entries 1, 3, 5, 8, 9 are yet to be classified. Recomputing the Gini for Status,

Gini index of **Department**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Sales** | **Systems** | **Marketing** |
| Junior | 1 | 0 | 1 |
| Senior | 1 | 1 | 1 |

Gini index of **Salary**:

|  |  |  |  |
| --- | --- | --- | --- |
| **Salary** | **Low** | **Medium** | **High** |
| Junior | 1 | 1 | 0 |
| Senior | 0 | 2 | 1 |

Summarizing,

|  |  |  |
| --- | --- | --- |
| **Attribute** | **Gini** | **Gain** |
| Department | 0.4 | 0.08 |
| Salary | 0.2667 | **0.2133** |

Gain(Salary) is greater among all the attributes. Thus, selected attribute for splitting is ‘Salary’.

C:\Users\foo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Q1_3.png

Now, only Department is remaining, thus adding it as a decision node to the tree and the final decision tree becomes as shown in the figure below:

C:\Users\foo\AppData\Local\Microsoft\Windows\INetCache\Content.Word\Q1_1.png

1. Compute the following accuracy on training data: (i) individual class accuracy (ii) overall class accuracy
2. Individual class accuracy:
3. Overall class accuracy:
4. For the following test data, predict the class label for each instance using the tree constructed in (a)

|  |  |  |  |
| --- | --- | --- | --- |
| **Department** | **Age** | **Salary** | **Status** |
| Sales | 21 – 30 | High | **Junior** |
| Systems | 21 – 30 | Medium | **Junior** |
| Marketing | 41 – 50 | High | **Senior** |
| Marketing | 31 – 40 | Low | **Junior** |

1. **Naive Bayes Classification**
2. State the assumption(s) made by Naive Bayes Classifier.

Naïve Bayes classifier assumes that the attributes are conditionally independent, given the class label. Mathematically it can be shown as follows:

1. Consider the following dataset given below in **Table 2.**

|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **C** | **Class** |
| 0 | 0 | 0 | + |
| 0 | 0 | 1 | - |
| 0 | 1 | 1 | - |
| 0 | 1 | 1 | - |
| 0 | 0 | 1 | + |
| 1 | 0 | 1 | + |
| 1 | 0 | 1 | - |
| 1 | 0 | 1 | - |
| 1 | 1 | 1 | + |
| 1 | 0 | 1 | + |

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. Predict class label when using the probabilities computed from (b).

Since, , the class label is

1. Estimate the following conditional probabilities using m-estimate approach, with p = 0.5, m = 4.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. Predict class label when using the probabilities computed from (d).

Since, , the class label is

1. Compare the two methods for estimating probabilities. Which method is better and why?

m-estimate approach is better. Because, if class conditional probability of any class is zero, then the overall posterior probability for the class becomes zero. In such cases, m-estimate takes care that the overall posterior probability for the class remains non-zero by adding term in the numerator and in the denominator.

1. **Holt’s 1-Rule Method**

Holt’s 1-Rule method is described as shown below:

For each attribute , form a rule as follows:

For each value from the domain of ,

Select the set of instances where has value .

Let be the most frequent class in that set.

Add the following clause to the rule for :

If has value , then the class is

Calculate the classification accuracy of this rule.

Use the rule with the highest classification accuracy.

1. Apply Holt’s 1-Rule for the following dataset. All attributes are categorical. Show the rules and accuracy for each attribute (A, B, C).

|  |  |  |  |
| --- | --- | --- | --- |
| **A** | **B** | **C** | **Class** |
| 0 | 0 | 1 | - |
| 0 | 0 | 0 | - |
| 0 | 1 | 0 | - |
| 0 | 1 | 0 | - |
| 0 | 0 | 1 | + |
| 1 | 0 | 1 | + |
| 1 | 0 | 1 | - |
| 1 | 0 | 1 | - |
| 1 | 1 | 0 | + |
| 1 | 0 | 0 | + |

**Attribute A:**

Domain of A is {0, 1}.

For  (Because, there are 4 tuples with and 1 tuple with )

For  (Because, there are 2 tuples with and 3 tuples with )

**Rule:**

**Accuracy:**

(Misclassified tuple 5, 7, and 8)

**Attribute B:**

Domain of B is {0, 1}.

For  (Because, there are 4 tuples with and 3 tuples with )

For  (Because, there are 2 tuples with and 1 tuples with )

**Rule:**

**Accuracy:**

(Misclassified tuple 5, 6, 9, and 10)

**Attribute C:**

Domain of C is {0, 1}.

For  (Because, there are 3 tuples with and 2 tuples with )

For  (Because, there are 3 tuples with and 2 tuples with )

**Rule:**

**Accuracy:**

(Misclassified tuple 5, 6, 9, and 10)

1. Name the best attribute (i.e., attribute for which the total error is minimum). If there are more than 1 attribute with same accuracy, name all of them.

The best attribute among A, B, and C is A. It has 0.7 accuracy (misclassifies only 3 tuples)