**Constructing a Decision Tree**

**~~~PhucCoi~~~**

Using ID3 algorithm, construct (by hand, show detail work) a decision tree for below dataset. Is your decision tree overfitting? Can you use Pruning technique to reduce the problem? Explain your work.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| mpg | cylinders | displacement | horsepower | weight | acceleration | modelyear | maker |
| good | 4 | low | low | low | high | 75to78 | asia |
| bad | 6 | medium | medium | medium | medium | 70to74 | america |
| bad | 4 | medium | medium | medium | low | 75to78 | europe |
| bad | 8 | high | high | high | low | 70to74 | america |
| bad | 6 | medium | medium | medium | medium | 70to74 | america |
| bad | 4 | low | medium | low | medium | 70to74 | asia |
| bad | 4 | low | medium | low | low | 70to74 | asia |
| bad | 8 | high | high | high | low | 75to78 | america |
| bad | 8 | high | high | high | low | 70to74 | america |
| good | 8 | high | medium | high | high | 79to83 | america |
| bad | 8 | high | high | high | low | 75to78 | america |
| good | 4 | low | low | low | low | 79to83 | america |
| bad | 6 | medium | medium | medium | high | 75to78 | america |
| good | 4 | medium | low | low | low | 79to83 | america |
| good | 4 | low | low | medium | high | 79to83 | america |
| bad | 8 | high | high | high | low | 70to74 | america |
| good | 4 | low | medium | low | medium | 75to78 | europe |
| bad | 5 | medium | medium | medium | medium | 75to78 | europe |

Predict Miles-per-gallon?

|  |  |  |  |
| --- | --- | --- | --- |
| Modelyear | 75to78 | 70to74 | 79to83 |
| GOOD | 2 | 0 | 4 |
| BAD | 5 | 7 | 0 |
|  | **7** | **7** | **4** |

**Using GINI Index:**

GINI(75to78) = 1-(2/7)2 - (5/7)2 = 0.408 GINI(70to74) = 1-(0/7)2-(7/7)2 = 0

GINI(79to83) = 1-(4/4)2-(0/4)2 = 0

**GINIS(Modelyear)**= (7/18)\*GINI(75to78)+(7/18)\*GINI(70to74)+(4/18)\* GINI(79to83) = 0.1587

|  |  |  |  |
| --- | --- | --- | --- |
| Maker | asia | america | europe |
| GOOD | 1 | 4 | 1 |
| BAD | 2 | 8 | 2 |
|  | **3** | **12** | **3** |

GINI(asia) = 1-(1/3)2 - (2/3)2 = 0.444 GINI(america) = 1-(4/12)2-(8/12)2 = 0.444

GINI(europe) = 1-(1/3)2-(2/3)2 = 0.444

**GINIS(Maker)**= (3/18)\*GINI(asia)+(12/18)\*GINI(america)+(3/18)\* GINI(europe) = 0.444

|  |  |  |  |
| --- | --- | --- | --- |
| Acceleration | high | medium | low |
| GOOD | 3 | 1 | 2 |
| BAD | 1 | 4 | 7 |
|  | **4** | **5** | **9** |

GINI(high) = 1-(3/4)2 - (1/4)2 = 0.375 GINI(medium) = 1-(1/5)2-(4/5)2 = 0.32

GINI(low) = 1-(2/9)2-(7/9)2 = 0.346

**GINIS(Acceleration)**= (4/18)\*GINI(high)+(5/18)\*GINI(medium)+(9/18)\* GINI(low) = 0.3452

|  |  |  |  |
| --- | --- | --- | --- |
| Weight | high | medium | low |
| GOOD | 1 | 1 | 4 |
| BAD | 5 | 5 | 2 |
|  | **6** | **6** | **6** |

GINI(high) = 1-(1/6)2 - (5/6)2 = 0.278 GINI(medium) = 1-(1/6)2-(5/6)2 = 0.278

GINI(low) = 1-(4/6)2-(2/6)2 = 0.444

**GINIS(Weight)**= (6/18)\*GINI(high)+( 6/18)\*GINI(medium)+( 6/18)\* GINI(low) = 0.332

|  |  |  |  |
| --- | --- | --- | --- |
| Hoursepower | high | medium | low |
| GOOD | 5 | 2 | 4 |
| BAD | 0 | 7 | 0 |
|  | **5** | **9** | **4** |

GINI(high) = 1-(5/5)2 - (0/5)2 = 0 GINI(medium) = 1-(2/9)2-(7/9)2 = 0.346

GINI(low) = 1-(4/4)2-(0/4)2 = 0

**GINIS(Hoursepower) =** (9/18)\*GINI(medium) = 0.173

|  |  |  |  |
| --- | --- | --- | --- |
| Displacement | high | medium | low |
| GOOD | 1 | 1 | 4 |
| BAD | 5 | 5 | 2 |
|  | **6** | **6** | **6** |

GINI(high) = 1-(1/6)2 - (5/6)2 = 0.278 GINI(medium) = 1-(1/6)2 - (5/6)2 = 0. 278

GINI(low) = 1-(4/6)2-(2/6)2 = 0.444

**GINIS(Displacement)**= (6/18)\*GINI(high)+( 6/18)\*GINI(medium)+( 6/18)\* GINI(low) = 0.332

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cylinders | 4 | 5 | 6 | 8 |
| GOOD | 5 | 1 | 0 | 1 |
| BAD | 3 | 0 | 3 | 5 |
|  | **8** | **1** | **3** | **6** |

GINI(4) = 1-(5/8)2 - (3/8)2 = 0.4686

GINI(5) = 1-(1/1)2-(0/1)2 = 0

GINI(6) = 1-(0/3)2-(3/3)2 = 0

GINI(8) = 1-(1/6)2-(5/6)2 = 0.2778

**GINIS(Cylinders)**= (8/18)\*GINI(4)+(6/18)\*GINI(8) = 0.30087

**GINIS(Modelyear)** = 0.1587 which is the smallest GINI value compared to other columns. Therefore, we are going to take **ModelYear** to classify

**70to74**

**Modelyear**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| mpg | cylinders | Displacement | hoursepower | weight | acceleration | maker |
| Bad | 6 | medium | medium | medium | medium | america |
| Bad | 8 | high | high | high | low | america |
| Bad | 6 | medium | medium | medium | medium | america |
| Bad | 4 | low | medium | low | medium | asia |
| Bad | 4 | low | hedium | low | low | asia |
| Bad | 8 | high | high | high | low | america |
| Bad | 8 | high | high | high | low | america |

(I)

d

**79to83**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| mpg | cylinders | Displacement | hoursepower | weight | acceleration | maker |
| good | 8 | high | medium | high | high | america |
| good | 4 | low | low | low | low | america |
| good | 4 | medium | low | low | low | america |
| good | 4 | low | low | medium | high | america |

(II)

**75to78**

(III)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| mpg | cylinders | Displacement | hoursepower | weight | acceleration | maker |
| good | 4 | low | low | low | high | asia |
| bad | 4 | medium | medium | medium | low | europe |
| bad | 8 | high | high | high | low | america |
| bad | 8 | high | high | high | low | america |
| bad | 6 | medium | medium | medium | high | america |
| good | 4 | low | medium | low | medium | europe |
| bad | 5 | medium | medium | medium | medium | europe |

Relying on table marked (I), we can totally conclude that if **Modelyear** is **70to74** which Miles-per-gallon definitelly is **good.**

Similarity table marked (I), table (II) also can conclude that if **Modelyear** is **79to83** which Miles-per-gallon definitelly is **bad.**

Now we move on to table marked (III) to classify it by continuously using GINI index. And by doing that we will get **displacement** column having the smallest GINI value.

**displacement**

**low**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| mpg | cylinders | hoursepower | weight | acceleration | maker |
| good | 4 | low | low | high | asia |
| good | 4 | medium | low | medium | europe |

(I)

**high**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| mpg | cylinders | hoursepower | weight | acceleration | maker |
| bad | 8 | high | high | low | america |
| bad | 8 | high | high | low | america |

(II)

**medium**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| mpg | cylinders | hoursepower | weight | acceleration | maker |
| bad | 4 | medium | medium | low | europe |
| bad | 6 | medium | medium | high | america |
| bad | 5 | medium | medium | medium | europe |

(III)

Basing on there marked table above which is really clear classification.

* In case **displacement** is **low** which Miles-per-gallon definitelly is **good.**
* In case **displacement** is **high** which Miles-per-gallon definitelly is **bad.**
* In case **displacement** is **medium** which Miles-per-gallon definitelly is **bad.**

Finally, we can draw a decision tree like this:

**70to74**

**Modelyear**

**75to78**

**79to83**

**Displacement**

**BAD**

**GOOD**

**medium**

**high**

**low**

**BAD**

**BAD**

**GOOD**

***Is your decision tree overfitting?***

To answer for the question whether or not my decision tree overfitting, let’s look at conditions when a decision tree occur overffiting:

• Irrelevant attributes can result in overfitting the training example data

* If hypothesis space has many dimensions (large number of attributes), we may find meaningless regularity in the data that is irrelevant to the true, important, distinguishing features
* About this, our dataset only have 7 features that affect the result of Mile per-gallon value, which I think having any irrelevant attributes in data.

• If we have too little training data, even a reasonable hypothesis space will‘overfit’

* Yes, it is might caus e our dataset is overfitting. So how to avoid overfiting in decision tree? By doing these regulations we might avoid overfiting:
  + Stop growing when data split is not statistically significant
  + Acquire more training data
  + Remove irrelevant attributes (manual process – not always possible)
  + Grow full tree, then post-prune

In conclusion, our dataset might be overfit by acquiring more training data. We can totally avoid overfiting.

***Can you use Pruning technique to reduce the problem?***

No, I don’t think we can use pruning technique to reduce the problem simply because our tree is just having 2 layers/levels which is short enough to don’t need to cutting back the tree