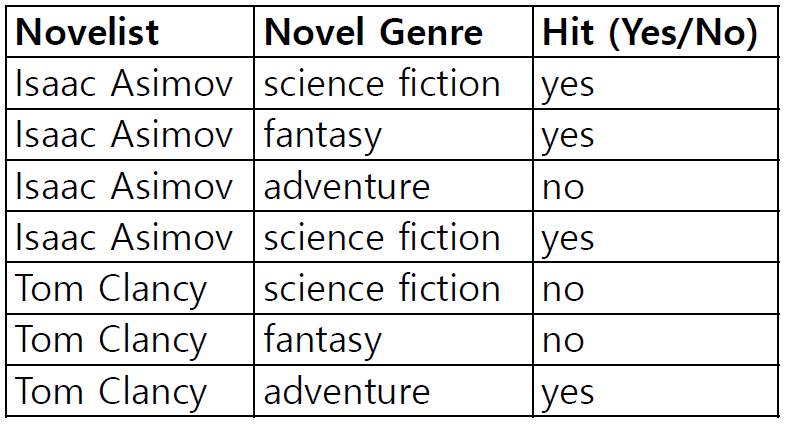
201533661 이승수’s Machine Learning Homework#2

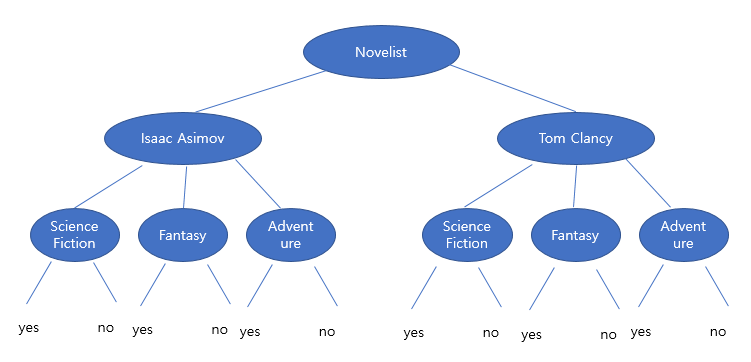
**Written Homework1**: Build a Decision Tree for the Following Dataset, Using Gini Index

Here is a new novel manuscript



We will choose the decision tree with smaller Gini Index at root node for classification.

**Case1: Gini Index of Decision Tree with Novelist index as a root node**



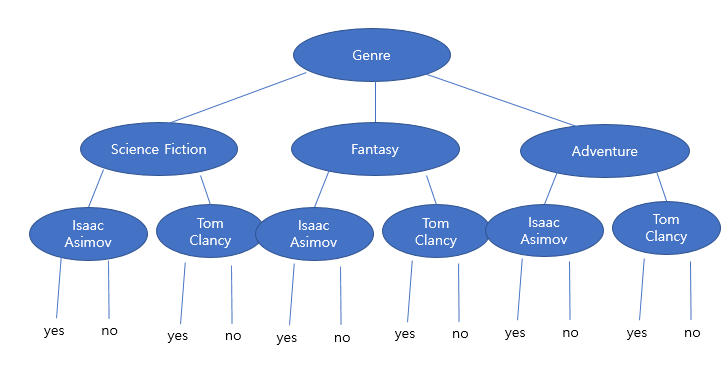
|  |  |  |
| --- | --- | --- |
|  | Isaac Asimov | Tom Clancy |
| Science Fiction | 2 | 1 |
| Fantasy | 1 | 1 |
| Adventure | 1 | 1 |

Gini(Isaac Asimov)=1-(1/2)^2-(1/4)^2\*2=5/8

Gini(Tom Clancy)=1-(1/3)^2\*3=2/3

Gini(Novelist)=(4/7)\*(5/8)+(3/7)\*(2/3)=9/14

**Case2: Gini Index of Decision Tree with Genre index as a root node**



|  |  |  |  |
| --- | --- | --- | --- |
|  | Science Fiction | Fantasy | Adventure |
| Isaac Asimov | 2 | 1 | 1 |
| Tom Clancy | 1 | 1 | 1 |

Gini(Science Fiction)=1-(1/2)^2-(1/4)^2\*2=5/8

Gini(Fantasy)=1-(1/2)^2\*2=1/2

Gini(Adventure)=1-(1/2)^2\*2=1/2

Gini(Genre)=(3/7)\*(4/9)+(2/7)\*(1/2)+(2/7)\*(1/2)=10/21

**Result: Gini Index of Genre node(10/21) is smaller than it of Novelist node(9/14), so we will choose dicision tree with Genre index as a root node.**

**Written Homework2**: Selecting the Best Splitting Value for a Continuous Variable



We will choose better method to choose splitting value among Naïve and Efficient Method.

**[Naive Method]**

**Case1)Efficient Method when choosing split position at Taxable Income=60**

|  |  |  |
| --- | --- | --- |
|  | 60<=TI | TI<60 |
| Yes | 0 | 3 |
| No | 1 | 6 |

Gini(60<TI)=1-0-1=0

Gini(TI<=60)= 1-(3/9)^2-(6/9)^2=4/9

Gini Index=(1/10)\* Gini(60<TI)+(9/10)\* Gini(TI<=60)= 2/5 =0.4

**Case2)Efficient Method when choosing split position at Taxable Income=70**

|  |  |  |
| --- | --- | --- |
|  | 70<=TI | TI<70 |
| Yes | 0 | 3 |
| No | 2 | 5 |

Gini(70<TI)=1-0-1=0

Gini(TI<=70)= 1-(3/8)^2-(5/8)^2=15/32

Gini Index=(2/10)\* Gini(TI<=70)+(8/10)\* Gini(TI<=70)= 3/8 =0.375

**Case4)Efficient Method when choosing split position at Taxable Income=75**

|  |  |  |
| --- | --- | --- |
|  | 75<=TI | TI<75 |
| Yes | 0 | 3 |
| No | 3 | 4 |

Gini(75<TI)=1-0—1=0

Gini(TI<=75)= 1-(3/7)^2-(4/7)^2=24/49

Gini Index=(3/10)\* Gini(75<TI)+(7/10)\* Gini(TI<=75)= 12/35 =0.343

**Case5)Efficient Method when choosing split position at Taxable Income=85**

|  |  |  |
| --- | --- | --- |
|  | 85<=TI | TI<85 |
| Yes | 1 | 2 |
| No | 4 | 3 |

Gini(85<TI)=1-(1/5)^2-(4/5)^2=8/25

Gini(TI<=85)= 1-(2/5)^2-(3/5)^2=12/25

Gini Index=(5/10)\* Gini(85<TI)+(5/10)\* Gini(TI<=85) =10/25 =0.4

**Case6)Efficient Method when choosing split position at Taxable Income=90**

|  |  |  |
| --- | --- | --- |
|  | 90<=TI | TI<90 |
| Yes | 2 | 1 |
| No | 3 | 4 |

Gini(90<TI)=1-(2/5)^2-(3/5)^2=12/25

Gini(TI<=90)= 1-(1/5)^2-(4/5)^2=8/25

Gini Index=(5/10)\* Gini(90<TI)+(5/10)\* Gini(TI<=90)= 2/5 =0.4

**Case7)Efficient Method when choosing split position at Taxable Income=95**

|  |  |  |
| --- | --- | --- |
|  | 95<=TI | TI<95 |
| Yes | 3 | 0 |
| No | 3 | 4 |

Gini(95<TI)=1-(3/6)^2-(3/6)^2=1/2

Gini(TI<=95)= 1-0-1=0

Gini Index=(6/10)\* Gini(95<TI) +(4/10)\* Gini(TI<=95)=0.3

**Case8)Efficient Method when choosing split position at Taxable Income=100**

|  |  |  |
| --- | --- | --- |
|  | 100<=TI | TI<100 |
| Yes | 3 | 0 |
| No | 4 | 3 |

Gini(100<TI)=1-(3/7)^2-(4/7)^2=24/49

Gini(TI<=100)= 1-0-1=0

Gini Index=(7/10)\* Gini(100<TI)+(3/10)\* Gini(TI<=100)= 12/35 =0.42

**Case9)Efficient Method when choosing split position at Taxable Income=120**

|  |  |  |
| --- | --- | --- |
|  | 120<=TI | TI<120 |
| Yes | 3 | 0 |
| No | 5 | 2 |

Gini(120<TI)=1-(3/8)^2-(5/8)^2=15/32

Gini(TI<=120)= 1-0-1=0

Gini Index=(8/10)\* Gini(120<TI)+(2/10)\* Gini(TI<=120)= 3/8 =0.375

**Case10)Efficient Method when choosing split position at Taxable Income=125**

|  |  |  |
| --- | --- | --- |
|  | 125<=TI | TI<125 |
| Yes | 3 | 0 |
| No | 6 | 1 |

Gini(125<TI)=1-(3/9)^2-(6/9)^2=4/9

Gini(TI<=125)= 1-0-1=0

Gini Index=(9/10)\* Gini(125<TI)+(1/10)\* Gini(TI<=125)= 2/5 =0.4

**Case11)Efficient Method when choosing split position at Taxable Income=220**

|  |  |  |
| --- | --- | --- |
|  | 220<=TI | TI<220 |
| Yes | 3 | 0 |
| No | 7 | 0 |

Gini(220<TI)=1-(3/10)^2-(7/10)^2=21/50

Gini(TI<=220)= 1-0-0=1

Gini Index=1\* Gini(220<TI)+0\* Gini(TI<=220)= 21/50 =0.42

->At Efficient Method, We will choose the split position between Taxable Income=95,100(case7) with the smallest Gini index with 0.3.

**[Efficient Method]**

**case1)Efficient Method when choosing split position before Taxable Income=60**

|  |  |  |
| --- | --- | --- |
|  | 55<=TI | TI<55 |
| Yes | 0 | 3 |
| No | 0 | 7 |

Choose split value smaller than 60=55

Gini(55<TI)=1-0-0=1

Gini(TI<=55)= 1-(3/10)^2-(7/10)^2=21/50

Gini Index=0\* Gini(55<TI)+1\* Gini(TI<=55)= 21/50 =0.42

**case2)Efficient Method when choosing split position between Taxable Income=60,70**

|  |  |  |
| --- | --- | --- |
|  | 65<=TI | TI<65 |
| Yes | 0 | 3 |
| No | 1 | 6 |

Median(60,70)=65

Gini(65<TI)=1-0-1=0

Gini(TI<=65)= 1-(3/9)^2-(6/9)^2=4/9

Gini Index=(1/10)\* Gini(65<TI)+(9/10)\* Gini(TI<=65)= 2/5 =0.4

**Case3)Efficient Method when choosing split position between Taxable Income=70,75**

|  |  |  |
| --- | --- | --- |
|  | 72<=TI | TI<72 |
| Yes | 0 | 3 |
| No | 2 | 5 |

Median(70,75)=72

Gini(72<TI)=1-0-1=0

Gini(TI<=72)= 1-(3/8)^2-(5/8)^2=15/32

Gini Index=(2/10)\* Gini(TI<=72)+(8/10)\* Gini(TI<=72)= 3/8 =0.375

**Case4)Efficient Method when choosing split position between Taxable Income=75,85**

|  |  |  |
| --- | --- | --- |
|  | 80<=TI | TI<80 |
| Yes | 0 | 3 |
| No | 3 | 4 |

Median(75,85)=80

Gini(80<TI)=1-0—1=0

Gini(TI<=55)= 1-(3/7)^2-(4/7)^2=24/49

Gini Index=(3/10)\* Gini(80<TI)+(7/10)\* Gini(TI<=55)= 12/35 =0.343

**Case5)Efficient Method when choosing split position between Taxable Income=85,90**

|  |  |  |
| --- | --- | --- |
|  | 87<=TI | TI<87 |
| Yes | 1 | 2 |
| No | 4 | 3 |

Median(85,90)=87

Gini(87<TI)=1-(1/5)^2-(4/5)^2=8/25

Gini(TI<=87)= 1-(2/5)^2-(3/5)^2=12/25

Gini Index=(5/10)\* Gini(87<TI)+(5/10)\* Gini(TI<=87) =10/25 =0.4

**Case6)Efficient Method when choosing split position between Taxable Income=90,95**

|  |  |  |
| --- | --- | --- |
|  | 92<=TI | TI<92 |
| Yes | 2 | 1 |
| No | 3 | 4 |

Median(90,95)=92

Gini(92<TI)=1-(2/5)^2-(3/5)^2=12/25

Gini(TI<=92)= 1-(1/5)^2-(4/5)^2=8/25

Gini Index=(5/10)\* Gini(92<TI)+(5/10)\* Gini(TI<=92)= 2/5 =0.4

**Case7)Efficient Method when choosing split position between Taxable Income=95,100**

|  |  |  |
| --- | --- | --- |
|  | 97<=TI | TI<97 |
| Yes | 3 | 0 |
| No | 3 | 4 |

Median(95,100)=97

Gini(97<TI)=1-(3/6)^2-(3/6)^2=1/2

Gini(TI<=55)= 1-0-1=0

Gini Index=(6/10)\* Gini(97<TI) +(4/10)\* Gini(TI<=55)=0.3

**Case8)Efficient Method when choosing split position between Taxable Income=100,120**

|  |  |  |
| --- | --- | --- |
|  | 110<=TI | TI<110 |
| Yes | 3 | 0 |
| No | 4 | 3 |

Median(100,120)=110

Gini(110<TI)=1-(3/7)^2-(4/7)^2=24/49

Gini(TI<=110)= 1-0-1=0

Gini Index=(7/10)\* Gini(110<TI)+(3/10)\* Gini(TI<=110)= 12/35 =0.42

**Case9)Efficient Method when choosing split position between Taxable Income=120,125**

|  |  |  |
| --- | --- | --- |
|  | 122<=TI | TI<122 |
| Yes | 3 | 0 |
| No | 5 | 2 |

Median(120,125)=122

Gini(122<TI)=1-(3/8)^2-(5/8)^2=15/32

Gini(TI<=122)= 1-0-1=0

Gini Index=(8/10)\* Gini(122<TI)+(2/10)\* Gini(TI<=122)= 3/8 =0.375

**Case10)Efficient Method when choosing split position between Taxable Income=125,220**

|  |  |  |
| --- | --- | --- |
|  | 175<=TI | TI<175 |
| Yes | 3 | 0 |
| No | 6 | 1 |

Median(125,220)=175

Gini(175<TI)=1-(3/9)^2-(6/9)^2=4/9

Gini(TI<=175)= 1-0-1=0

Gini Index=(9/10)\* Gini(175<TI)+(1/10)\* Gini(TI<=175)= 2/5 =0.4

**Case11)Efficient Method when choosing split position bigger than Taxable Income=220**

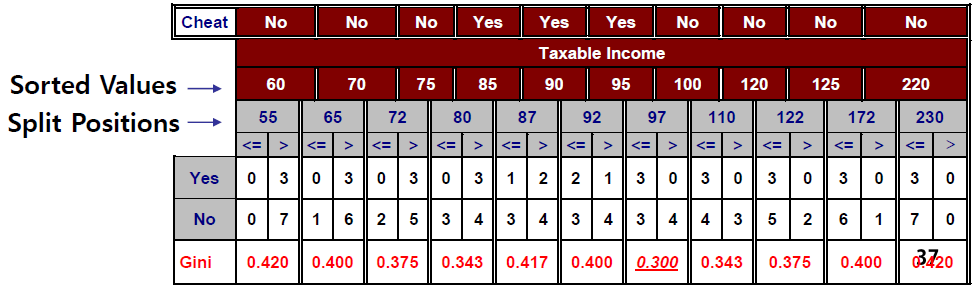
|  |  |  |
| --- | --- | --- |
|  | 230<=TI | TI<230 |
| Yes | 3 | 0 |
| No | 7 | 0 |

Split position bigger than 220=230

Gini(230<TI)=1-(3/10)^2-(7/10)^2=21/50

Gini(TI<=230)= 1-0-0=1

Gini Index=1\* Gini(230<TI)+0\* Gini(TI<=230)= 21/50 =0.42



->At Efficient Method, We will choose the split position between Taxable Income=95,100(case7) with the smallest Gini index with 0.3.

**Result: We believe that Efficient Method is the best way to choose splitting value for Taxable Income because we can make more elaborate cases than Naïve Method.**