Assignment 2

1.

a. Gini Index for all samples

GINIsplit =

b. Gini index for the Customer ID attribute

*is the same at each node*

GINIsplit

c. Gini index for the Gender Attribute

*M*

*F*

GINI­split =

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

*Family*

*Sports*

*Luxury*

GINIsplit =

e. Gini Index for the shirt-size

*small*

*medium*

*large*

*extra large*

GINIsplit =

f. It follows, then, Shirt Size is a better split than Gender and Car Type.

g. The 20-way customer-ID split stands as a unproductive split because it provides no correlation between attributes and class-labels. The split is fairly arbitrary and no relationship can be adequately gleaned between the nature of the split and the applicable class labels.

2.

Recall: Entropy =

InfoGain = -

splitinfo =

GainRatio = InfoGain/split­­info

Entropyparent =

All possible splits:

Customer ID, Gender, CarType, Shirt Size

Customer ID -

|  |  |  |
| --- | --- | --- |
| Record | Class 0 | Class 1 |
| 1 | 1 | 0 |
| 2 | 1 | 0 |

... it continues in a similar pattern

Entropy for each =

InfoGain =

splitinfo =

GainRatio = 1.0 / splitinfo = http://www4b.wolframalpha.com/Calculate/MSP/MSP71541a5aie6dg73e519b00003f6d7c784b56gf61?MSPStoreType=image/gif&s=62&w=443&h=20

Gender -

|  |  |  |
| --- | --- | --- |
| Record | C0 | C1 |
| M | 6 | 4 |
| F | 4 | 6 |

Entropy for *M* =

Entropy for *F* = http://www4b.wolframalpha.com/Calculate/MSP/MSP29061a5b0077f04306h300005cga1c83f913i1he?MSPStoreType=image/gif&s=49&w=443&h=20

(It's the same as the *Male* split, because the class proportions are the same

InfoGain = http://www4b.wolframalpha.com/Calculate/MSP/MSP29061a5b0077f04306h300005cga1c83f913i1he?MSPStoreType=image/gif&s=49&w=443&h=20=

http://www5a.wolframalpha.com/Calculate/MSP/MSP19861a5b00ffai616aa3000014d3794gbhd08fhg?MSPStoreType=image/gif&s=52&w=443&h=20

splitinfo  = = 1.0

GainRatio = http://www5a.wolframalpha.com/Calculate/MSP/MSP19861a5b00ffai616aa3000014d3794gbhd08fhg?MSPStoreType=image/gif&s=52&w=443&h=20

Car Type

3 way split

|  |  |  |
| --- | --- | --- |
| Record | C0 | C1 |
| Family | 1 | 3 |
| Sports | 8 | 0 |
| Luxury | 1 | 7 |

3way

*Family* Entropy =

= http://www5a.wolframalpha.com/Calculate/MSP/MSP20691a5b0206a1625ih100004322bfhe016d05bi?MSPStoreType=image/gif&s=58&w=443&h=20

*Sports* Entropy =

*Luxury* Entropy =

InfoGain =

- =http://www4a.wolframalpha.com/Calculate/MSP/MSP134031a5ahfa7ad7e3cad000062fed8h119h0d775?MSPStoreType=image/gif&s=18&w=75&h=20

splitinfo = = http://www5a.wolframalpha.com/Calculate/MSP/MSP13941a5b04f3i156iai800005ga03df540ic2g12?MSPStoreType=image/gif&s=63&w=443&h=20

GainRatio = http://www5a.wolframalpha.com/Calculate/MSP/MSP1001a5b0532aeg1h56g0000351b0e60ge397e60?MSPStoreType=image/gif&s=53&w=75&h=20

2 way split - 1

|  |  |  |
| --- | --- | --- |
| Record | C0 | C1 |
| Family/Sports | 9 | 3 |
| Luxury | 1 | 7 |

*I think the car-type is almost ordinal and this association is a little strange... that is, to put family and sports or family and luxury together is a little strange. I'll let them be.*

2 way split - 2

|  |  |  |
| --- | --- | --- |
| Record | C0 | C1 |
| Family/Luxury | 2 | 10 |
| Sports | 8 | 0 |

*see above*

2 way split - 3

|  |  |  |
| --- | --- | --- |
| Record | C0 | C1 |
| Family | 1 | 3 |
| Sports/Luxury | 9 | 7 |

*Family* Entropy = http://www5a.wolframalpha.com/Calculate/MSP/MSP20691a5b0206a1625ih100004322bfhe016d05bi?MSPStoreType=image/gif&s=58&w=443&h=20

*Sports/Luxury* Entropy =

= http://www5a.wolframalpha.com/Calculate/MSP/MSP4181a5b05f2285dc7de000049fgd4062a0cch76?MSPStoreType=image/gif&s=45&w=443&h=20

InfoGAIN = http://www5a.wolframalpha.com/Calculate/MSP/MSP45611a5aiigeg8a02fea00004bb248dfg8dd1dhc?MSPStoreType=image/gif&s=35&w=83&h=20

splitINFO = = http://www5a.wolframalpha.com/Calculate/MSP/MSP6521a5b059iccg90hhb0000335h626ac166e777?MSPStoreType=image/gif&s=58&w=443&h=20

GainRATIO = http://www5a.wolframalpha.com/Calculate/MSP/MSP65151a5aihigf71eh73i0000330ig412880he7ai?MSPStoreType=image/gif&s=55&w=83&h=20

Shirt Size

4 way split

|  |  |  |
| --- | --- | --- |
| Record | C0 | C1 |
| Small | 3 | 2 |
| Medium | 3 | 4 |
| Large | 2 | 2 |
| X-Large | 2 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 0.970951 | 0.203509 | 1.958872 | 0.103891 |
| 0.985228 |  |  |  |
| 1 |  |  |  |
| 1 |  |  |  |

2 way split - 1

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Small/Medium | 6 | 6 |
| Large/X-Large | 4 | 4 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 1 | -0.00892 | 0.970951 | -0.00919 |
| 1 |  |  |  |

2 way split - 2

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Small/Medium/Large | 8 | 8 |
| X-Large | 2 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 1 | -0.00892 | 0.721928 | -0.01236 |
| 1 |  |  |  |

2 way split - 3

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Small | 3 | 2 |
| Medium/Large/X-Large | 7 | 8 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 0.970951 | 0.000745 | 0.811278 | 0.000918 |
| 0.996792 |  |  |  |

3 way split - 1

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Small/Medium | 6 | 6 |
| Large | 2 | 2 |
| X-Large | 2 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 0.00892 | -1.370951 | -0.00651 |
| 1 |  |  |  |
| 1 |  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Small | 3 | 2 |
| Medium/Large | 5 | 6 |
| X-Large | 2 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 0.970951 | 0.001622 | 1.438759 | 0.001127 |
| 0.99403 |  |  |  |
| 1 |  |  |  |

|  |  |  |
| --- | --- | --- |
|  |  |  |
| Small | 3 | 2 |
| Medium | 3 | 4 |
| Large/X-Large | 4 | 4 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 0.970951 | 0.003509 | 1.558872 | 0.002251 |
| 0.985228 |  |  |  |
| 1 |  |  |  |

(At first, I was spelling out the values and using a calculator... I moved to excel after a while... I moved kind of quickly because there were so many to do...)

It seems that the best split is the three way car split.

3.

a. Entropy with respect to positive class

http://www4a.wolframalpha.com/Calculate/MSP/MSP6581a5b0807c4f6i3d900004a957ghgd5f58i23?MSPStoreType=image/gif&s=40&w=443&h=20

b. infoGains of a1 and a2

a1

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

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 0.811278 | 0.229437 | 0.991076 | 0.231503 |
| 0.721928 |  |  |  |

a2

|  |  |  |
| --- | --- | --- |
|  | + | - |
| T | 2 | 3 |
| F | 2 | 2 |

|  |  |  |  |
| --- | --- | --- | --- |
| Entropy | InfoGain | splitINFO | GainRatio |
| 0.970951 | 0.007215 | 0.991076 | 0.00728 |
| 1 |  |  |  |

**c.**

For the split at -4.0 the following analyses apply:

GINIIndices: 1.0 for <= -4.0 / 0.49382716049382713 for > -4.0.

GINISplit: 0.49382716049382713

Entropy: 0.0 for <= -4.0 / 0.9910760598382222 for > -4.0.

InfoGain: 0.0.

GainRatio: 0.0.

For the split at 2.0 the following analyses apply:

GINIIndices: 0.0 for <= 2.0 / 0.46875 for > 2.0.

GINISplit: 0.41666666666666663

Entropy: -0.0 for <= 2.0 / 0.9544340029249649 for > 2.0.

InfoGain: 0.14269027946047563.

GainRatio: 0.28353286890735246.

For the split at 3.5 the following analyses apply:

GINIIndices: 0.5 for <= 3.5 / 0.48979591836734704 for > 3.5.

GINISplit: 0.49206349206349215

Entropy: 1.0 for <= 3.5 / 0.9852281360342514 for > 3.5.

InfoGain: 0.002565287367137792.

GainRatio: 0.003356807432159333.

For the split at 4.5 the following analyses apply:

GINIIndices: 0.4444444444444444 for <= 4.5 / 0.4444444444444444 for > 4.5.

GINISplit: 0.4444444444444444

Entropy: 0.9182958340544896 for <= 4.5 / 0.9182958340544896 for > 4.5.

InfoGain: 0.07278022578373267.

GainRatio: 0.07925575079916354.

For the split at 5.5 the following analyses apply:

GINIIndices: 0.48 for <= 5.5 / 0.5 for > 5.5.

GINISplit: 0.4888888888888889

Entropy: 0.9709505944546687 for <= 5.5 / 1.0 for > 5.5.

InfoGain: 0.007214618474517431.

GainRatio: 0.007279581019942209.

For the split at 6.5 the following analyses apply:

GINIIndices: 0.5 for <= 6.5 / 0.4444444444444444 for > 6.5.

GINISplit: 0.48148148148148145

Entropy: 1.0 for <= 6.5 / 0.9182958340544896 for > 6.5.

InfoGain: 0.018310781820059074.

GainRatio: 0.01993995958711118.

For the split at 7.5 the following analyses apply:

GINIIndices: 0.5 for <= 7.5 / 0.0 for > 7.5.

GINISplit: 0.4444444444444444

Entropy: 1.0 for <= 7.5 / -0.0 for > 7.5.

InfoGain: 0.10218717094933338.

GainRatio: 0.20305112481622145.

For the split at 13.0 the following analyses apply:

GINIIndices: 0.49382716049382713 for <= 13.0 / 1.0 for > 13.0.

GINISplit: 0.49382716049382713

Entropy: 0.9910760598382222 for <= 13.0 / 0.0 for > 13.0.

InfoGain: 0.0.

GainRatio: 0.0.

It looks like the best split is on a2 with an infogain > 0.2

f.

A1 A2

|  |  |  |  |
| --- | --- | --- | --- |
| GINI index | GINISPlit | GINI index | GINISPlit |
| 0.375 | 0.344444 | 0.48 | 0.488889 |
| 0.32 |  | 0.5 |  |

it seems that a1 is a better split than a2, according to GINI.

4. (a program.)

5.

Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | YES (predicted) | NO (predicted) |
| YES(Actual) | 2 (TP) | 3 (FN) |
| NO(Actual) | 5 (FP) | 5 (TN) |

