MDL A-4

Tanvi Karandikar
2018101059

Data Used

Data I am working with is dataset 10 with entries 3 and 6 flipped:

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
2	Yes	Good	No	No
3	Yes	Average	Yes	No
4	Yes	Average	No	No
5	Yes	Bad	Yes	Yes
6	Yes	Bad	No	No
7	No	Good	Yes	No
8	No	Good	No	No
9	No	Average	Yes	Yes
10	No	Average	No	No
11	No	Bad	Yes	Yes
12	No	Bad	No	No

Method Used for Calculation

(The method described in the textbook has been followed)

Formula for entropy of a node is:

$$B(q) = -(q \log_2 q + (1-q) \log_2 (1-q))$$

Formula for Information Gain:

$$Gain(A) = B(\frac{p}{p+n}) - Remainder(A)$$

where

An attribute A with d distinct values divides the training set E into subsets $E_1, \ldots, E_d$. Each subset E_k has p_k positive examples and n_k negative examples, so if we go along that branch, we will need an additional $B(p_k/(p_k + n_k))$ bits of information to answer the question. A randomly chosen example from the training set has the kth value for the attribute with probability $(p_k + n_k)/(p + n)$, so the expected entropy remaining after testing attribute A is

$$Remainder(A) = \sum_{k=1}^{d} \frac{p_k+n_k}{p+n} B(\frac{p_k}{p_k+n_k})$$

Note: B(0) is taken as 0 ie when all the entries have the same value entropy is 0.

Finding the Tree

Step 1

Data I am working with:

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
2	Yes	Good	No	No
3	Yes	Average	Yes	No
4	Yes	Average	No	No
5	Yes	Bad	Yes	Yes
6	Yes	Bad	No	No
7	No	Good	Yes	No
8	No	Good	No	No
9	No	Average	Yes	Yes
10	No	Average	No	No
11	No	Bad	Yes	Yes
12	No	Bad	No	No

Entropy at this stage:
Number of **Yes** = 4/12
Number of **No** = 8/12

$B(q) = -q\log(q) - (1 - q)\log(1 - q)$
=> $B(4/12) = -(4/12)\log(4/12) + -(8/12)\log(8/12) =$ **0.918296**

Three ways we can split the data:

1. Single

Node Single=Yes (6/12 nodes)

Num	**Single**	**Performance**	**CAG**	**Solved**
1	Yes	Good	Yes	Yes
2	Yes	Good	No	No
3	Yes	Average	Yes	No
4	Yes	Average	No	No
5	Yes	Bad	Yes	Yes
6	Yes	Bad	No	No

Entropy of this Node:
Number of **Yes** = 2/6
Number of **No** = 4/6

B(2/6) = **0.918296**

Node Single=No (6/12 nodes)

Num	**Single**	**Performance**	**CAG**	**Solved**
7	No	Good	Yes	No
8	No	Good	No	No
9	No	Average	Yes	Yes
10	No	Average	No	No
11	No	Bad	Yes	Yes
12	No	Bad	No	No

Entropy of this Node:
Number of **Yes** = 2/6
Number of **No** = 4/6

B(2/6) = **0.918296**

So,
Gain(Single) = 0.918296+((6/12)*0.918296+(6/12)*0.918296)= 0

2. Performance

Node Performance=Good(4/12)

Num	**Single**	**Performance**	**CAG**	**Solved**
1	Yes	Good	Yes	Yes
2	Yes	Good	No	No
7	No	Good	Yes	No
8	No	Good	No	No

Entropy of this Node:
Number of **Yes** = 1/4
Number of **No** = 3/4

B(1/4) = **0.811278**

Node Performance=Average(4/12)

Num	**Single**	**Performance**	**CAG**	**Solved**
3	Yes	Average	Yes	No
4	Yes	Average	No	No
9	No	Average	Yes	Yes
10	No	Average	No	No

Entropy of this Node:
Number of **Yes** = 1/4
Number of **No** = 3/4

B(1/4) = **0.811278**

Node Performance=Bad(4/12)

Num	Single	Performance	CAG	Solved
5	Yes	Bad	Yes	Yes
6	Yes	Bad	No	No
11	No	Bad	Yes	Yes
12	No	Bad	No	No

Entropy of this Node:
Number of **Yes** = 2/4
Number of **No** = 2/4

B(2/4) = **1**

So,
Gain(Performance) = 0.918296-((4/12)*0.811278+(4/12)*0.811278+(4/12)*1)= **0.044110667**

3. CAG

Node CAG=Yes(6/12)

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
3	Yes	Average	Yes	No
5	Yes	Bad	Yes	Yes
7	No	Good	Yes	No
9	No	Average	Yes	Yes
11	No	Bad	Yes	Yes

Entropy of this Node:
Number of **Yes** = 4/6
Number of **No** = 2/6

B(4/6) = **0.918296**

Node CAG=No(6/12)

Num	**Single**	**Performance**	**CAG**	**Solved**
2	Yes	Good	No	No
4	Yes	Average	No	No
6	Yes	Bad	No	No
8	No	Good	No	No
10	No	Average	No	No
12	No	Bad	No	No

Entropy of this Node:
Number of **Yes** = 0/6
Number of **No** = 6/6

B(6/6) = **0**

So,
Gain(CAG) = 0.918296-((6/12)*0.918296+(6/12)*0) = **0.459148**

Now, comparing Gain in all the cases:
Gain(Single)=0
Gain(Performance)=0.044110667
Gain(CAG)=0.459148

So, we will use **CAG** to form the next nodes for the tree.

Step 1.1

(node where CAG=Yes)

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
3	Yes	Average	Yes	No
5	Yes	Bad	Yes	Yes
7	No	Good	Yes	No
9	No	Average	Yes	Yes
11	No	Bad	Yes	Yes

Entropy of this Node:
Number of **Yes** = 4/6
Number of **No** = 2/6

B(4/6) = **0.918296**

Two ways we can split the data:

1. Single

Node Single=Yes (3/6)

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
3	Yes	Average	Yes	No
5	Yes	Bad	Yes	Yes

Entropy of this Node:
Number of **Yes** = 2/3
Number of **No** = 1/3

B(2/3) = **0.918296**

Node Single=No (3/6)

Num	Single	Performance	CAG	Solved
7	No	Good	Yes	No
9	No	Average	Yes	Yes
11	No	Bad	Yes	Yes

Entropy of this Node:
Number of **Yes** = 2/3
Number of **No** = 1/3

B(2/3) = **0.918296**

So,
Gain(Single) = 0.918296-((3/6)*0.918296+(3/6)*0.918296) = **0**

2. Performance

Node Performance=Good (2/6)

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
7	No	Good	Yes	No

Entropy of this Node:
Number of **Yes** = 1/2
Number of **No** = 1/2

B(1/2) = **1**

Node Performance=Average (2/6)

Num	Single	Performance	CAG	Solved
3	Yes	Average	Yes	No
9	No	Average	Yes	Yes

Entropy of this Node:
Number of **Yes** = 1/2
Number of **No** = 1/2

B(1/2) = **1**

Node Performance=Bad (2/6)

Num	**Single**	**Performance**	**CAG**	**Solved**
5	Yes	Bad	Yes	Yes
11	No	Bad	Yes	Yes

Entropy of this Node:
Number of **Yes** = 2/2
Number of **No** = 0/2

B(2/2) = **0**

So,
Gain(Performance) = 0.918296-((2/6)*1+(2/6)*1+(2/6)*0) = **0.251629333**

Now, comparing Gain in all the cases:
Gain(Single)=0
Gain(Performance)=0.251629333

So, we will use **Performance** to form the next nodes for the tree.

Step 1.2

(node where CAG=No)

Node CAG=No(6/12)

Num	Single	Performance	CAG	Solved
2	Yes	Good	No	No
4	Yes	Average	No	No
6	Yes	Bad	No	No
8	No	Good	No	No
10	No	Average	No	No
12	No	Bad	No	No

Entropy of this Node:
Number of **Yes** = 0/6
Number of **No** = 6/6

B(6/6) = **0**

All the entries having CAG=No have their Solved=No.
So this node is a deterministic condition and we need not create any sub-nodes at this point, i.e. this is a leaf node.

Step 1.1.1

(node where CAG=Yes and Performance=Good)

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes
7	No	Good	Yes	No

Entropy of this Node:
Number of **Yes** = 1/2
Number of **No** = 1/2

B(1/2) = **1**

Only one attribute is left to split the data and that is Single

Single

Node Single=Yes(½)

Num	**Single**	**Performance**	**CAG**	**Solved**
1	Yes	Good	Yes	Yes

Entropy of this Node:
Number of **Yes** = 1/1
Number of **No** = 0/1

B(1/1)=0

Node Single=No(½)

Num	**Single**	**Performance**	**CAG**	**Solved**
7	No	Good	Yes	No

Entropy of this Node:
Number of **Yes** = 0/1
Number of **No** = 1/1

B(0)=0

So,
Gain(Single) = 1 - ((1/2)*0+(1/2)*0) = **1**

Since there is only one possible split this will be split nodes by **Single** attribute.

Step 1.1.2

(node where CAG=Yes and Performance=Average)

Num	Single	Performance	CAG	Solved
3	Yes	Average	Yes	No
9	No	Average	Yes	Yes

Entropy of this Node:
Number of **Yes** = 1/2
Number of **No** = 1/2

B(1/2) = **1**

Only one attribute is left to split the data and that is Single

Single

Node Single=No(½)

Num	Single	Performance	CAG	Solved
9	No	Average	Yes	Yes

Entropy of this Node:
Number of **Yes** = 1/1
Number of **No** = 0/1

B(0/1)=0

Node Single=Yes(½)

Num	Single	Performance	CAG	Solved
3	Yes	Average	Yes	No

Entropy of this Node:
Number of **Yes** = 0/1
Number of **No** = 1/1

B(0)=0

So,
Gain(Single) = 1 - ((1/2)*0+(1/2)*0) = **1**

Since there is only one possible split this will be split nodes by **Single** attribute.

Step 1.1.3

(node where CAG=Yes and Performance=Bad)

Num	**Single**	**Performance**	**CAG**	**Solved**
5	Yes	Bad	Yes	Yes
11	No	Bad	Yes	Yes

Entropy of this Node:
Number of **Yes** = 2/2
Number of **No** = 0/2

B(2/2) = **0**

All the entries having CAG=Yes and Performance=Bad have their Solved=No.
So this node is a deterministic condition and we need not create any sub-nodes at this point, i.e. this is a leaf node.

Step 1.1.1.1

(node where CAG=Yes and Performance=Good and Single=Yes)

Num	Single	Performance	CAG	Solved
1	Yes	Good	Yes	Yes

Entropy of this Node:
Number of **Yes** = 1/1
Number of **No** = 0/1

B(1/1)=0

There is only one entry here, and it's value is Yes.
So this node is a deterministic condition and we need not create any sub-nodes at this point, i.e. this is a leaf node.

Step 1.1.1.2

(node where CAG=Yes and Performance=Good and Single=No)

Num	Single	Performance	CAG	Solved
7	No	Good	Yes	No

Entropy of this Node:
Number of **Yes** = 0/1
Number of **No** = 1/1

B(0)=0

There is only one entry here, and it's value is No.
So this node is a deterministic condition and we need not create any sub-nodes at this point, i.e. this is a leaf node.

Step 1.1.2.1

(node where CAG=Yes and Performance=Average and Single=No)

Num	Single	Performance	CAG	Solved
9	No	Average	Yes	Yes

Entropy of this Node:
Number of **Yes** = 1/1
Number of **No** = 0/1

B(0/1)=0

There is only one entry here, and it's value is Yes.
So this node is a deterministic condition and we need not create any sub-nodes at this point, i.e. this is a leaf node.

Step 1.1.2.2

(node where CAG=Yes and Performance=Average and Single=Yes)

Num	Single	Performance	CAG	Solved
3	Yes	Average	Yes	No

Entropy of this Node:
Number of **Yes** = 0/1
Number of **No** = 1/1

B(0)=0

There is only one entry here, and it's value is No.
So this node is a deterministic condition and we need not create any sub-nodes at this point, i.e. this is a leaf node.

The Final Tree:

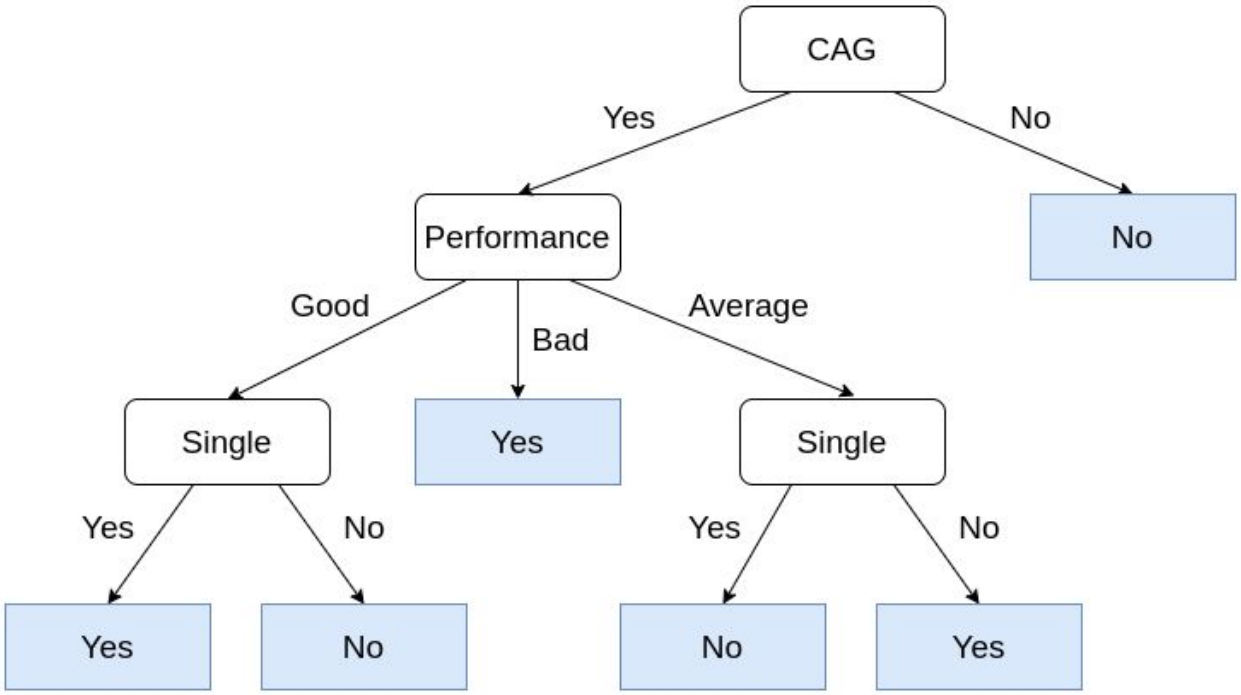
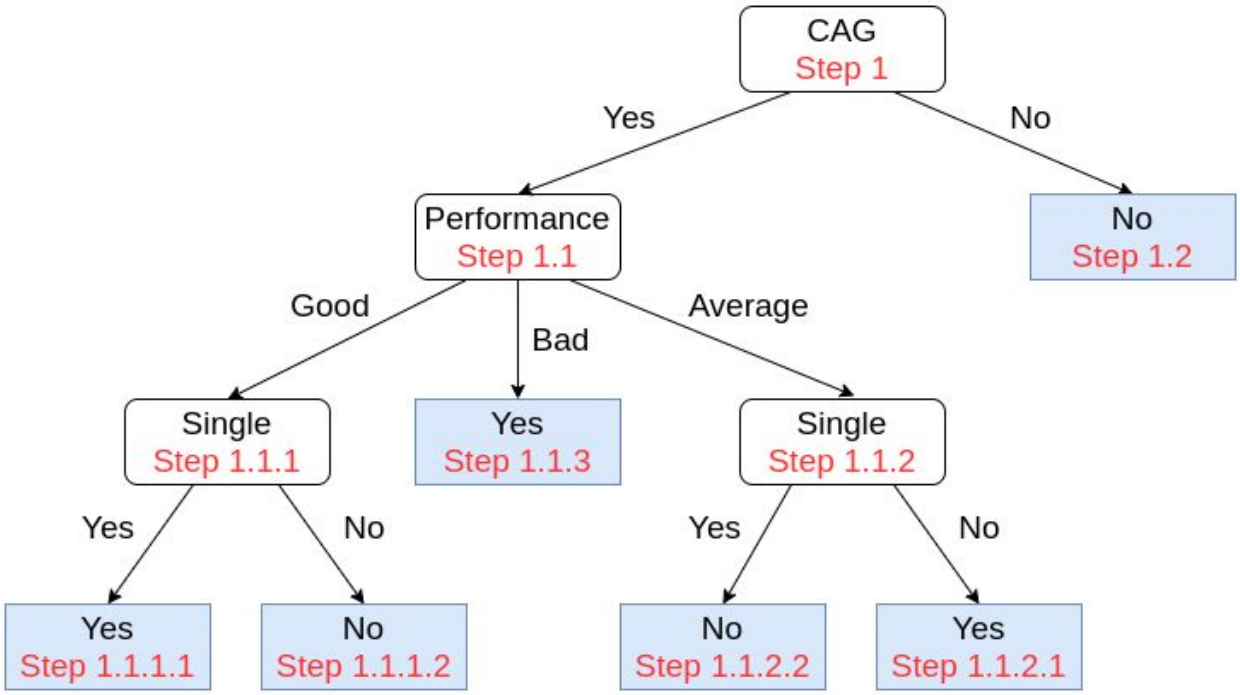


Diagram with numbering of steps:



Explanation of why a particular node was chosen along with calculation of the entropy and information gain can be viewed in the corresponding step number labelled.