Machine Data and Learning-Assignment 4

Anubhav Sharma - 2018114007

Question to do → Question 8

Following is the question table after modification:

Country	Flu/Cough	SFT	Positive Case
China	Yes	Yes	No
China	Yes	No	Yes
China	No	Yes	Yes
Italy	Yes	No	No
Italy	Yes	Yes	Yes
Italy	No	Yes	Yes
India	Yes	Yes	Yes
India	Yes	No	Yes
India	No	No	No
USA	No	Yes	Yes
USA	Yes	No	No

Theory :

Following are the concepts, formulas and notations used in this report→

- $E = -p(+) * log(p(+)) - p(-) * log(p(-))$

  - Here p(-) refers to the negative booleans or the number of options in negative for the event and vice versa for p(+)
  - This formula is analogous to :B(q) = −(q log 2 q + (1 − q) log 2 (1 − q))
- $Avg\ E\ =\ n1/n * E1\ +\ n2/n * E2\ +\ n3/n * E3\ ....$
- $Gain(E)\ = E(parent)\ -\ E(Avg\ E)$

where Ei is the entropy of the ith child of the node we want to calculate Avg E for.ni is the number of entries in the ith child and n is the total entries in parent.

Solution:

Level 0 choice:

E(Parent) → Here the parent is the entire class attribute or the entire output from the training set.

So

n(No)=4 , n(Yes)=7

So E(Parent)= 0.94566

F[0] → Country

Avg E(Country)= (3/11 * E(China)) + (3/11 * E(Italy)) + (3/11 * E(India)) + (2/11 * E(USA))

- ➔ E(China) ⇒ n(No)=1 , n(Yes)=2
  - ◆ E(China)=0.918295
- ➔ E(Italy) ⇒ n(No)=1 , n(Yes)=2
  - ◆ E(Italy)=0.918295
- ➔ E(India) ⇒ n(No)=1 , n(Yes)=2
  - ◆ E(India)=0.918295
- ➔ E(USA) ⇒ n(No)=1 , n(Yes)=1
  - ◆ E(USA)=1

Gain =0.0125

F[1] → Flu/Cough

_Avg E(Flu/Cough)= (7/11 * E(Yes)) + (4/11 * E(No))

- ➔ E(Yes) ⇒ n(No)=3 , n(Yes)=4
  - ◆ E(Yes)=0.98523
- ➔ E(No) ⇒ n(No)=1 , n(Yes)=3
  - ◆ E(No)=0.811278

Gain =0.023689

F[2] → SFT

Avg E(SFT)= (6/11 * E(Yes)) + (5/11 * E(No))

- ➔ E(Yes) ⇒ n(No)=1 , n(Yes)=5
  - ◆ E(Yes)=0.65002
- ➔ E(No) ⇒ n(No)=3 , n(Yes)=2
  - ◆ E(No)=0.970951

Gain=0.149762

Conclusion→ Now we have ***SFT*** with the maximum gain so it will become the root node now and its children will act as the parents for the further levels.

So splitting of the parameter of SFT

Level 1 choice(SKF):

Parents → Yes and No

SO when Parent is Yes →

Level 1 → SubLevel 0 ⇒ choice(SKF="Yes"):

E(Parent) →E(Yes)=0.65002

F[0] → Country

Avg E(Country)(given → SKF="YES")= (2/6 * E(China)) + (2/6 * E(Italy)) + (1/6 * E(India)) + (1/6 * E(USA))

- ➔ E(China) ⇒ n(No)=1 , n(Yes)=1
  - ◆ E(China)=1
- ➔ E(Italy) ⇒ n(No)=0 , n(Yes)=2
  - ◆ E(Italy)=0
- ➔ E(India) ⇒ n(No)=0 , n(Yes)=1
  - ◆ E(India)=0
- ➔ E(USA) ⇒ n(No)=0 , n(Yes)=1
  - ◆ E(USA)=0

Gain=0.31668

F[1] → Flu/Cough

_Avg E(Flu/Cough)(given → SKF="YES")= (3/6 * E(Yes)) + (3/6 * E(No))

- ➔ E(Yes) ⇒ n(No)=1 , n(Yes)=2
  - ◆ E(Yes)=0.918295
- ➔ E(No) ⇒ n(No)=0 , n(Yes)=3
  - ◆ E(No)=0

Gain =0.1908725

Conclusion→ Now we have ***Country*** with the maximum gain so it will become the root node now and its children will act as the parents for the further levels. And the sub level is still the **Yes** Of SFT.

Level 1 → SubLevel 0 → Level 2 ⇒ Choice (country):

Parents which are available are China , Italy , India and the USA.

Except China, all others are pure, so no more splitting,and we select the only valid parent which is China.

E(Parent) = E(China) = 1

F[1] → Flu/Cough

Avg E(Flu/Cough)(given → SKF=”YES” and Country=China)= (1/2 * E(Yes)) + (1/2 * E(No))

- ➔ E(Yes) ⇒ n(No)=1 , n(Yes)=0
  - ◆ E(Yes)=0
- ➔ E(No) ⇒ n(No)=0 , n(Yes)=1
  - ◆ E(No)=0

Gain =1

Conclusion-This is the final level since all the nodes are pure for this sublevel. Now we move towards the next sublevel i.e *No for SFT.*

NOW when Parent is No →

Level 1 → SubLevel 1 choice(SKF=”No”):

E(Parent) →E(No)=0.970951

F[0] → Country

Avg E(Country)(given → SKF=”No”)= (1/5 * E(China)) + (1/5 * E(Italy)) + (2/5 * E(India)) + (1/5 * E(USA))

- ➔ E(China) ⇒ n(No)=0 , n(Yes)=1
  - ◆ E(China)=0
- ➔ E(Italy) ⇒ n(No)=1 , n(Yes)=0

- ◆ E(Italy)=0

- ➔ E(India) ⇒ n(No)=1 , n(Yes)=1
  - ◆ E(India)=1
- ➔ E(USA) ⇒ n(No)=1 , n(Yes)=0
  - ◆ E(USA)=0

Gain=0.570951

F[1] → Flu/Cough

_Avg E(Flu/Cough)(given → SKF="No")= (4/5 * E(Yes)) + (1/5 * E(No))

- ➔ E(Yes) ⇒ n(No)=2 , n(Yes)=2
  - ◆ E(Yes)=1
- ➔ E(No) ⇒ n(No)=1 , n(Yes)=0
  - ◆ E(No)=0

Gain=0.170951

Conclusion→ Now we have ***Country*** with the maximum gain so it will become the root node now and its children will act as the parents for the further levels. And the sub level is still the **No** Of SFT.

Level 1 → SubLevel 1 → Level 2 choice(Country):

Parents which are available are China , Italy , India and the USA.

Except India, all others are pure, so no more splitting,and we select the only valid parent which is India.

E(Parent) = E(India) = 1

F[1] → Flu/Cough

Avg E(Flu/Cough)(given → SKF=”No” and Country=India)= (1/2 * E(Yes)) + (1/2 * E(No))

- ➔ E(Yes) ⇒ n(No)=0 , n(Yes)=1
    - ◆ E(Yes)=0
- ➔ E(No) ⇒ n(No)=1 , n(Yes)=0
    - ◆ E(No)=0

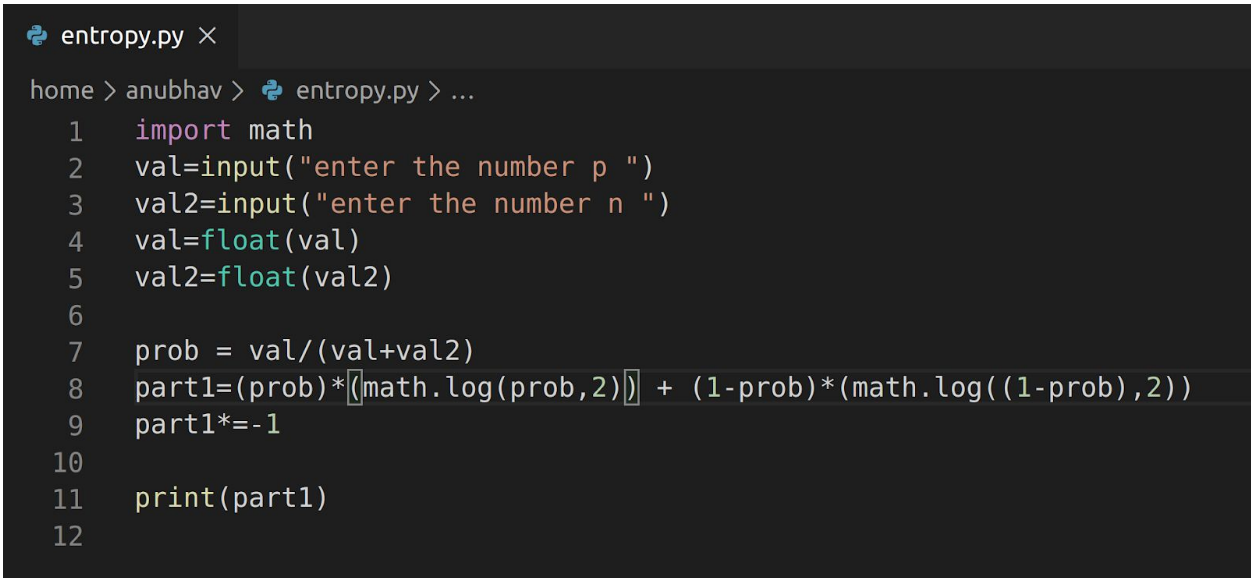
Gain =1

Max Gain is also 1. But we have reached the terminal stage.

Conclusion - Now all nodes are pure thus this is the final level.

So Our decision tree will have 3 levels.

Following is the script used for calculating entropy.



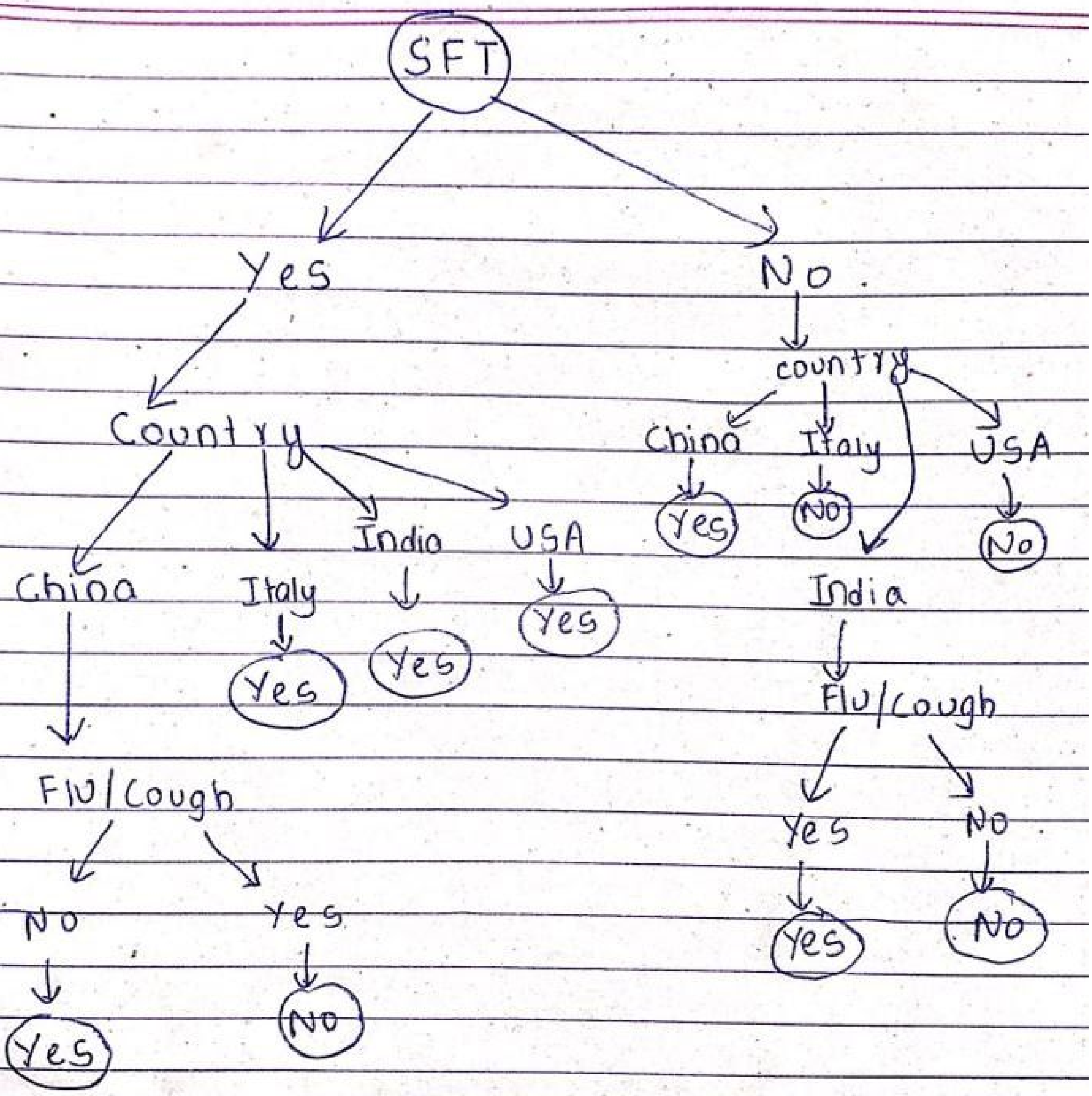
```python
import math
val=input("enter the number p ")
val2=input("enter the number n ")
val=float(val)
val2=float(val2)

prob = val/(val+val2)
part1=(prob)*(math.log(prob,2)) + (1-prob)*(math.log((1-prob),2))
part1*=-1

print(part1)

```

Decision TREE →


SFT
Yes
No
Country
country
China
Italy
India
USA
China
Italy
USA
Yes
No
No
Yes
Yes
Yes
India
Flu/Cough
Flu/Cough
Yes
No
No
Yes
Yes
No
Yes
No