1st step: Calculate the original entropy H(S)

Total = 6

H(S) → { 4 yes, 2 no }

$H(S) = -\frac{4}{6}\log_2(4/6) - \frac{2}{6}\log_2(2/6)$

$= 0.38 + 0.52 = 0.9$

chestpain:

yes : {3 yes, 0 no} Total = 3

No : {1 yes, 2 no} Total = 3

$Gain(S, chestpain) = H(S) - \frac{3}{6}H(yes) - \frac{3}{6}H(no)$ (3 yes, 3 no)

$H(yes, chestpain) = -\frac{3}{3}\log_2(3/3) - \frac{0}{3}\log_2(0/3) = 0$

$H(no, chestpain) = -\frac{1}{3}\log_2(1/3) - \frac{2}{3}\log_2(2/3) =$

$= 0.528 + 0.389 = 0.917$

$Gain(S, chestpain) = 0.9 - \frac{3}{6} \times 0 - \frac{3}{6} \times 0.917$

$= 0.9 - 0 - 0.4585$

$= 0.4415$

Again, Diet → 4 yes 2 no : (total)

yes : {2 yes, 2 no} Total = 4

no : {2 yes, 0 no} Total = 2

Gain(S, Diet) = $H(S) - \frac{4}{6} H(yes) - \frac{2}{6} H(no)$

$H(yes, Diet) = -\frac{2}{4}\log_2(2/4) - \frac{2}{4}\log_2(2/4) = 0.5 + 0.5 = 1$

$H(NO, Diet) = -\frac{2}{2}\log_2(2/2) - 0/2 \log_2(0/2)$

↓

0.

Gain(S, Diet) = $0.9 - \frac{4}{6} \times 1 - 2/6 \times 0$

$= 0.9 - 0.667 - 0$

$= 0.233$.

Also,

smoke : (4 yes, 2 no) In total

yes : {3 yes, 1 no} total = 4

no : {1 no, 0 yes} {1 yes, 1 no} total = 2

Gain(S, smoke) = $H(S) - \frac{4}{6} H(yes) - \frac{2}{6} H(no)$

$H(yes, smoke) = -\frac{3}{4}\log_2(3/4) - \frac{1}{4}\log_2(1/4)$

$= 0.311 + 0.5 = 0.811$

$H(no, smoke) = -\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2)$

$= 0.5 + 0.5 = 1$

Gain(S, smoke) = $0.9 - 4/6 \times 0.811 - 2/6 \times 1$

$0.9 - 0.540 - 0.33$ $= 0.03$

Exercise (4 yes, 2 no) in total

yes : {2 yes, 2 no} total = 4

no : {2 yes, 0 no} total = 2

$Gain(S, exercise) = H(S) - \frac{4}{6} H(yes) - \frac{2}{6} H(no)$

$H(yes, exercise) : -\frac{2}{4}\log_2(2/4) - \frac{2}{4}\log_2(2/4) = 1$

$H(no, exercise) : -\frac{2}{2}\log_2(2/2) - \frac{0}{2}\log_2(0/2) = 0.$

$Gain(S, exercise) = 0.9 - 4/6 \times 1 - 2/6 \times 0$

$= 0.9 - 0.666$

$= 0.233$

Compare all gain & choose the one with max gain. So, we consides chestpain as our root node.

{1, 2 ..., 6}

Chestpain [4 yes, 2 No]

yes

{1, 2, 5} → posn where yes is present

[3 yes, 0 No] → total = 3.

So, it is yes

no.

{3, 4, 6} posn where no is present in chestpain

[1 yes, 2 no] → total 3.

? ?

So, we need to calculate for chestpain → no.
consides only 3, 4 & 6.

ID	chest pain	diet	smoke	exercises	H.A
3	no	no	yes	no	yes
4	no	yes	no	yes	no
6	no	yes	yes	yes	no

[1 yes & 2 no of HA]

But we don't consider chest pain so, our table is like this:

ID	diet	smoke	exercises	HA
3	no	yes	no	yes
4	yes	no	yes	no
6	yes	yes	yes	no

Again, find gain of all 3 attributes.

diet (2 yes, 1 no) in total

yes: {0 yes, 2 no} total = 2.

no: {1 yes, 0 no} total = 1

~~Gain(S, diet) = H(S)~~

$Gain(S, diet) = H(S) - \frac{2}{3} H(yes) - \frac{1}{3} H(no)$

$H(yes, diet) = -\frac{0}{2}\log_2(0/2) - \frac{2}{2}\log_2(2/2) = 0$

Also,

$H(S) = -\frac{1}{3}\log_2(1/3) - \frac{2}{3}\log_2(2/3)$

$= 0.528 + 0.389$

$= 0.91$

so, Gain $= 0.91 - 0 = (0.91)$

Smoke (2 yes, 1 no) in total

yes: {1 yes, 1 no} 2 total

no: {1 no, 0 yes} 1 total

$Gain(S, smoke) = H(S) - \frac{2}{3} H(yes) - \frac{1}{3} H(no)$

$H(yes, smoke) = -\frac{1}{2}\log_2(\frac{1}{2}) - \frac{1}{2}\log_2(\frac{1}{2}) = 0.5 + 0.5 = 1$

$H(no, smoke) = -\frac{1}{1}\log_2(\frac{1}{1}) - \frac{0}{1}\log_2(\frac{0}{1}) = 0$

$Gain(S, smoke) = 0.91 - \frac{2}{3} \times 1 - 0$

$= 0.243$

Exercise (2 yes, 1 no) in total

yes: {0 yes, 2 no} 2 total

no: {1 yes, 0 no} 1 total

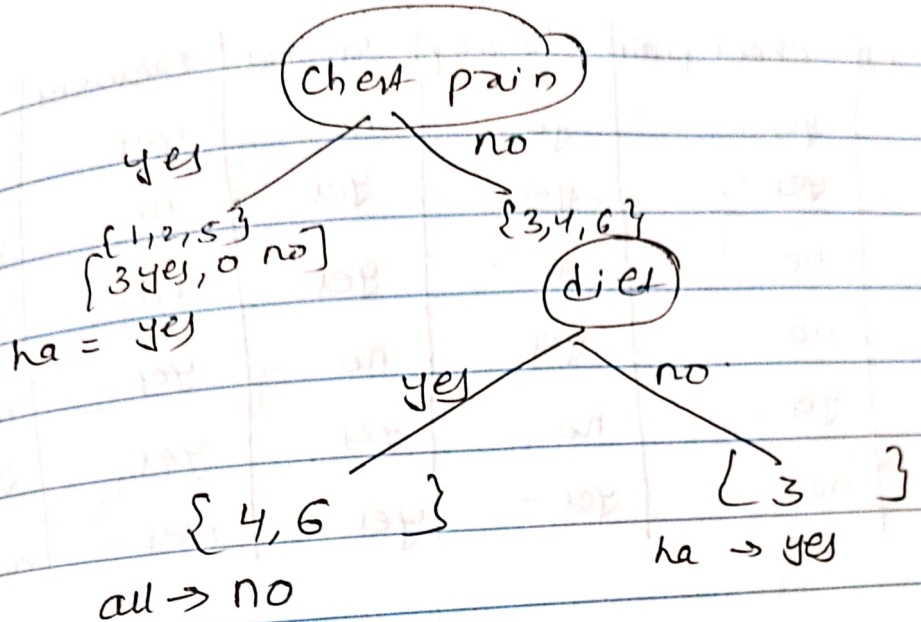
$Gain(S, exe) = H(S) - \frac{2}{3} H(yes) - \frac{1}{3} H(no)$

$H(yes, exercise) = -\frac{0}{2}\log_2(\frac{0}{2}) - \frac{2}{2}\log_2(\frac{2}{2}) = 0$

$H(no, exercise) = -\frac{1}{1}\log_2(\frac{1}{1}) - \frac{0}{1}\log_2(\frac{1}{1}) = 0$

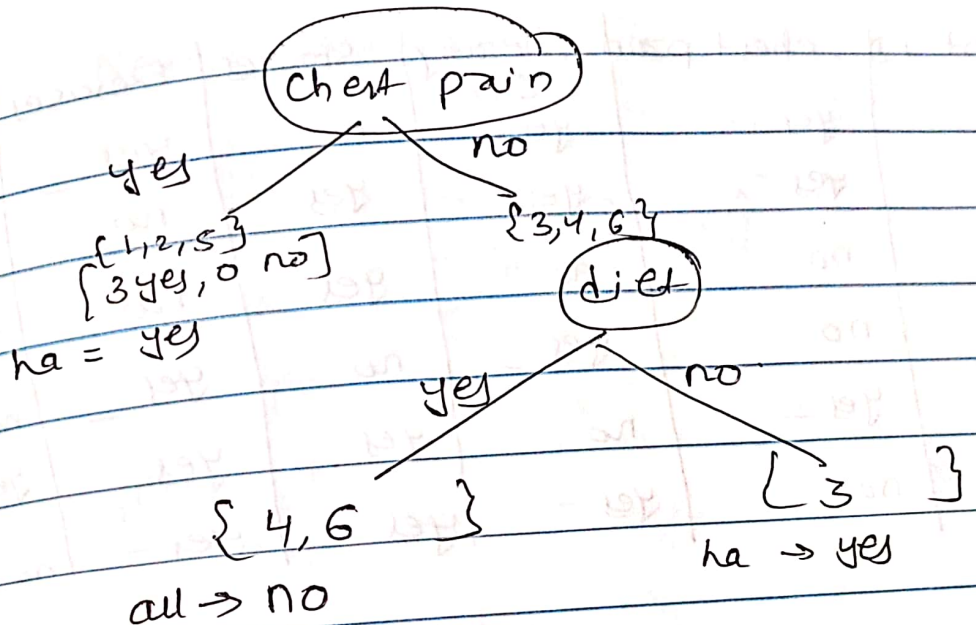
So, Gain = 0.91

Here, as diet & exercise have highest gain, choosing diet as our next root node.



Each branch ends in a homogeneous set of examples so the construction of decision tree ends here.

Use the above dataset and adaboost with T=2 to train a classifier. Use the decision stump as the weak learner. Report the value of α_1 and α_2 and decision stump f_1 & f_2.



Each branch ends in a homogeneous set of examples so the construction of decision tree ends here.

i) Use the above dataset and adaboost with T=2 to train a classifier. Use the decision stump as the weak learner. Report the value of α_1 and α_2 and decision stump f_1 & f_2.

Patient ID	chest pain	healthy diet	smoker	Exercises	Heart attack
1	yes	yes	no	yes	yes
2	yes	yes	yes	no	yes
3	no	no	yes	no	yes
4	no	yes	no	yes	no
5	yes	no	yes	yes	yes
6	no	yes	yes	yes	no

→ Here,

1st step: creating Decision stump for each of the attributes:

(a) chest pain

yes: 3 yes, 0 no

no: 1 yes, 2 no.

(one misclassification here)

(b) Healthy diet

yes: 2 yes, 2 no

no: 2 yes, 0 no

(4 misclassification here).

(c) Smokes

yes: 3 yes, 1 no

no: 1 yes, 1 no

(2 misclassification)

(d) Exercises

yes: (2 yes, 2 no)

no: (2 yes, 0 no)

(4 misclassification here)

Entropy is given by:

$$E = -P_y \ln(P_y) - P_n \ln(P_n)$$

For (a): yes: $-\frac{3}{3}\ln(\frac{3}{3}) - \frac{0}{3}\ln(\frac{0}{3}) = 0$

no: $-\frac{1}{3}\ln(\frac{1}{3}) - \frac{2}{3}\ln(\frac{2}{3}) = 0.63$

Entropy $= \frac{3}{6} \times 0 + \frac{3}{6} \times 0.63$

$= 0.315$

For: (b) yes: $-\frac{2}{4}\ln\left(\frac{2}{4}\right) - \frac{2}{4}\ln\left(\frac{2}{4}\right) = 0.693$

no: $\frac{-2}{2}\ln\left(\frac{2}{2}\right) - \frac{0}{2}\ln\left(\frac{0}{2}\right) = 0$

Entropy = $\frac{4}{6} \times 0.693 + \frac{2}{6} \times 0 = 0.462$

For: (c) yes: $-\frac{3}{4}\ln(3/4) - \frac{1}{4}\ln(1/4) = 0.56$

no: $-\frac{1}{2}\ln(1/2) - \frac{1}{2}\ln(1/2) = 0.69$

entropy = $4/6 \times 0.56 + \frac{2}{6} \times 0.69 = 0.60$

For (d) yes: $-\frac{2}{4}\ln(2/4) - 2/4\ln(2/4) = 0.693$

no: $-\frac{2}{2}\ln(2/2) - \frac{0}{2}\ln(0/2) = 0$

entropy = $4/6 \times 0.693 + 2/6 \times 0 = 0.462$.

Here, Entropy with least value:

entropy of chestpain = 0.315

Hence, In 1st pass, our weak learner is chestpain

So,

chestpain

yes — 3 yes, 0 no

no — 1 yes, 2 no (1 misclassification).

$$\text{weight}(\alpha_1) = \frac{1}{2}\ln\ln\left(\frac{1-\text{error}}{\text{error}}\right)$$

$$= \frac{1}{2}\ln\left(\frac{1-1/6}{1/6}\right)$$

Here, error = 1/6, because we only have one misclassification here:

Hence, $\alpha_1 = 0.80$

adjusted weight

	Chest Pain				H.A	sampleweight		Normalized weight
1	yes	yes	no	yes	yes	1/6	0.07	0.09
2	yes	yes	yes	no	yes	1/6	0.07	0.09
3	no	no	yes	no	yes	1/6	0.37	0.51
4	no	yes	no	yes	no	1/6	0.07	0.09
5	yes	no	yes	yes	yes	1/6	0.07	0.09
6	no	yes	yes	yes	no	1/6	0.07	0.09

New weight = $1/6 \times e^{0.80} = 0.37$

Also, to decrease weight: $1/6 \times e^{-0.80} = 0.07$ } Formula for Adjusted weight

3rd row is our misclassification here for chest pain.

Next step is to normalize weight:

Sum of Adjust weight values = 0.72

To Normalize: $\frac{0.07}{0.72} = 0.09$.

Normalized weights are the weights for 2nd step

So, our table is like this:

	chest pain	Healthy diet	smoke	exercise	H.A	sampleweight
1	yes	yes	no	yes	yes	0.09
2	yes	yes	yes	no	yes	0.09
3	no	no	yes	no	yes	0.51
4	no	yes	no	yes	no	0.09
5	yes	no	yes	yes	yes	0.09
6	no	yes	yes	yes	no	0.09

Our Dataset looks like:

	Chest Pain				H.A	Sample wt.	adjusted weight	Normalized wt
1	yes	yes	no	yes	yes	1/6	0.07	0.09
2	yes	yes	yes	no	yes	1/6	0.07	0.09
3	no	no	yes	no	yes	1/6	0.37	0.51
4	no	yes	no	yes	no	1/6	0.07	0.09
5	yes	no	yes	yes	yes	1/6	0.07	0.09
6	no	yes	yes	yes	no	1/6	0.07	0.09

New weight = $1/6 \times e^{0.80} = 0.37$

Also, to decrease weight: $1/6 \times e^{-0.80} = 0.07$ } Formula for Adjusted wt

3rd row is our misclassification here. for Stump Chest pain.

Next step is to normalize weight:

Sum of Adjust weight values = 0.72

To Normalize: $\frac{0.07}{0.72} = \underline{0.09}$.

Normalized weights are the weights for 2nd step

So, our table is like this:

	Chest Pain	Healthy diet	smoke	exercise	H.A	sample weight
1	yes	yes	no	yes	yes	0.09
2	yes	yes	yes	no	yes	0.09
3	no	no	yes	no	yes	0.51
4	no	yes	no	yes	no	0.09
5	yes	no	yes	yes	yes	0.09
6	no	yes	yes	yes	no	0.09

We create new Dataset with exactly 6 rows:

	C.P	H.D	smoke	Exer.	H.A	Bins:
1	no	yes	yes	yes	no	0-0.09
2	no	~~yes~~ no	~~no~~ yes	~~yes~~ no	~~no~~ yes	0.09-0.18
3	no	~~yes~~ no	~~no~~ yes	~~yes~~ no	~~no~~ yes	0.18-0.69
4	~~no~~ no	yes	~~no~~ no	yes	~~yes~~ no	0.69-0.78
5	~~yes~~ yes	yes	~~no~~ yes	~~yes~~ no	yes	0.78-0.87
6	no	yes	yes	yes	no	0.87-0.96

How to fill:

1. Change the weights to put into bins/range.
2. Generate random no: betⁿ 0 & 1

Let's say 1st is 0.93 & see it falls in which bin.

0.64 ✓

0.39 ✓

0.13 ✓

0.5 ✓

0.87

Create new Dataset with exactly 6 rows

	C.P	H.D	smoke	Exes	H.A	
1	no	yes	yes	yes	no	
2	no	no	yes	no	yes	repeat 3 times
3	no	no	yes	no	yes	
4	yes	yes	yes	no	yes	
5	no	no	yes	no	yes	
6	no	yes	yes	yes	no	

To our new dataset, the weightage of misclassi points is high.

Again, Repeating steps:

(a) chest pain

yes (4 yes, 0 no) — no (3 yes, 2 no)

(b) Healthy died

yes (4 yes, 2 no) — no (3 yes, 0 no)

(c) Smokes

yes (4 yes, 2 no) — no (2 yes, 0

Exercises

yes (~~4 yes, 2 no~~) — no (~~3 yes, 0 no~~)

(~~2 no~~, 0 yes, 2 no) (4 yes, 0 no)

(zero)

misclassification

here:

Then,

For chest pain: (yes) $= \frac{-1}{1} \ln(1/1) - \frac{0}{1} \ln(0/1) = 0$

(no) $= -3/5 \ln(3/5) - \left(\frac{2}{5}\right) \ln(2/5) = 067$

entropy $= \frac{1}{6} \times 0 + \frac{5}{6} \times 0.67 = 0.55$

For Healthy diet :

yes : $-\frac{1}{3}\ln(1/3) - \frac{2}{3}\ln(2/3) = 0.63$

no : $\frac{-3}{3}\ln(3/3) - \frac{0}{3}\ln(0/3) = 0$

entropy = $3/6 \times 0.63 + 0 = 0.315$

For Smoke

yes : $-4/6\ln(4/6) - 2/6\ln(2/6) = 0.63$

no : 0

entropy = $\frac{6}{6} \times 0.63 = 0.63$

For exercise

yes : $-\frac{0}{2}\ln(0/2) - \frac{2}{2}\ln(2/2) = 0$

no : $-4/4\ln(4/4) - 0/4\ln(0/4) = 0$

entropy = 0

Here, choosing entropy of exercise = 0

& so, our next weak learner is Exercise.

~~$\alpha_2 = \frac{1}{2}\ln($~~

Error for exercise is : 0.96

Then, $\alpha_2 = 1/2\ln\left(\frac{1-0.96}{0.96}\right)$

$= 1/2\ln(0.04)$

$= 0.$

$\alpha_2 = 1/2\ln\left(\frac{1-1.53}{1.53}\right)$

$0.51 \times 3 = 1.53$

$\ln\left(\frac{-0.53}{1.53}\right)$

$= 1/2\ln(-0.34)$

$= 0$