6)

Working of ADABOOST with Example.

Consider

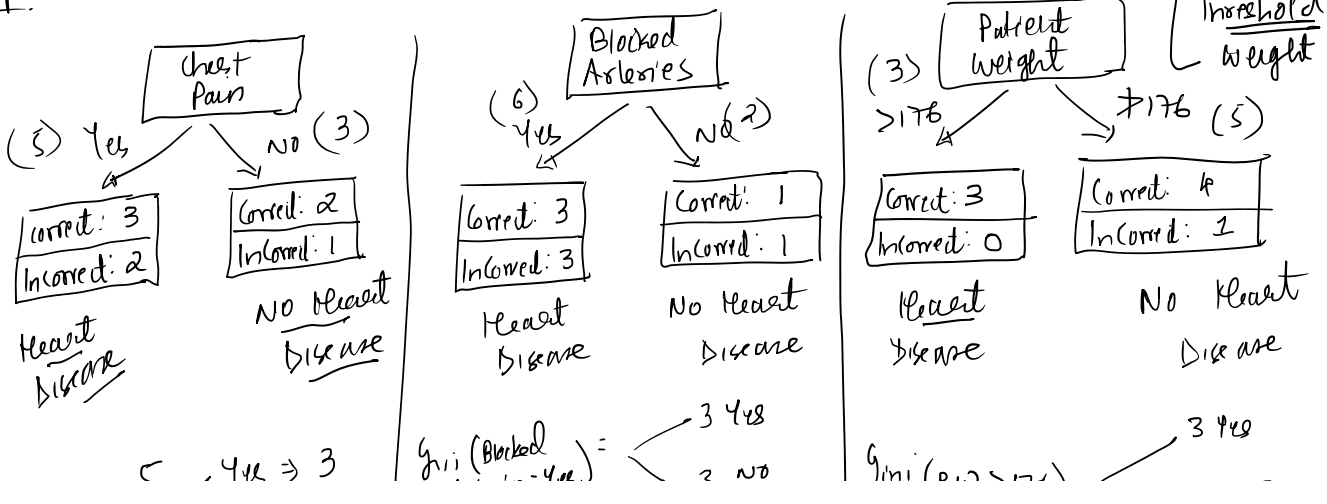
Chest Pain	Blocked Arteries	Patient weight	Heart Disease
Yes	Yes	205	Yes
NO	Yes	180	Yes
Yes	NO	210	Yes
Yes	Yes	167	Yes
NO	Yes	156	NO
NO	Yes	125	NO
Yes	NO	168	NO
Yes	Yes	172	NO

Step 1: Assign weight to every sample (Initial) $= \frac{1}{\text{Total No of Sample}} = \frac{1}{8}$

*(Initially Equal Weights are assigned to each Sample)

Chest Pain	Blocked Arteries	Patient weight	Heart Disease	Sample weight
Yes	Yes	205	Yes	1/8
NO	Yes	180	Yes	1/8
Yes	NO	210	Yes	1/8
Yes	Yes	167	Yes	1/8
NO	Yes	156	NO	1/8
NO	Yes	125	NO	1/8
Yes	NO	168	NO	1/8
Yes	Yes	172	NO	1/8

Step 2: Create Stump on Each Feature



Div:

Gini (Chest Pain = Yes) = 5 → Yes ⇒ 3, No ⇒ 2 (Heart Disease)

$= 1 - \left(\left(\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right)$

$= 0.48$

Gini (Chest Pain = No) = 3 → Yes = 2, No = 1

$= 1 - \left(\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right)$

$= 0.444$

Weighted Gini (Chest Pain)

$= 0.48 * \frac{5}{8} + 0.44 * \frac{3}{8}$

$= 0.4665 \approx 0.47$

Gini (Blocked Arteries = Yes) = 3 Yes, 3 No

$= 1 - \left(\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right)$

$= 0.5$

Gini (Blocked arteries = No) = 1 Yes, 1 No

$= 1 - \left(\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right)$

$= 0.5$

Gini (Blocked Arteries) = 0.5

Gini (P.W > 176) = 3 Yes, 0 No

$= 1 - \left(\left(\frac{3}{3}\right)^2\right)$

$= 0$

Gini (P.W ≯ 176) = 4 Yes, 1 No

$= 1 - \left(\left(\frac{4}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right)$

$= 0.32$

Gini (Weight) = 0.2

* The smallest Gini Index is 0.2 for weight >176 feature

So first Stump will be weight > 176

Step 3 =>

To calculate Amount of Say of weight > 176 [feature].

$$\text{Amount of say} = \frac{1}{2} \ln\left(\frac{1 - \text{Total Error}}{\text{Total Error}}\right)$$

Total Error: for a stump the Sum of weights associated with Incorrectly classified sample.

Here only one sample is Incorrectly classified.

Yes	Yes	167	Yes	1/8

∴ Total Error = 1/8

* Amt of say = $\frac{1}{2} \ln\left(\frac{1-(1/8)}{(1/8)}\right) = \frac{0.97}{(\text{High})}$

[Note : If Amt of say is less then the stump is not doing good job]

* Now we have → First stump (weight > 176)
→ Amt of say.

Now Initially Every Sample had Equal weight →

NOW Before Second stump we must modify weights of Samples

* Weights of Correctly classified Samples ⇒ must be decrease
" " Incorrectly classified Samples ⇒ must be increase
[So that next stump will focus more on Incorrectly identified Samples].

*

New Sample Weight (Correctly classified Sample) = Sample weight (old) * $e^{-\text{amount of say}}$

New Sample Weight (Incorrectly classified Sample) = Sample weight (old) * $e^{+\text{amount of say}}$

∴ for Incorrectly Classified Sample = $\frac{1}{8} * e^{+0.97} = 0.33$

For Correctly Classified Sample = $\frac{1}{8} * e^{-0.97} = 0.05$

Chest Pain ✓	Blocked Arteries ✓	Patient weight ✓	Heart Disease ✓	Sample weight	New Sample weight ↓	Normalize Sample weight ↓
Yes.	Yes	205	Yes.	1/8	0.05	0.05/0.68 = 0.07
NO.	Yes	180	Yes.	1/8	0.05	0.07
Yes.	No	210	Yes.	1/8	0.05	0.07
Yes.	Yes	167	Yes	1/8	0.33	0.33/0.68 = 0.49

Yes.	No	210	Yes.	1/8	0.05	
Yes.	Yes	167	Yes	1/8	0.33	0.33/0.68 = 0.49
No	Yes	156	No	1/8	0.05	0.07
No.	Yes	125	No.	1/8	0.05	0.07
Yes.	No	168	No	1/8	0.05	0.07
Yes	Yes	172	No	1/8	0.05	0.07

↑ Sum = 0.68

* Now we have New Normalized Sample weights.

* The Incorrectly classified Sample has Higher New Sample weight
  → which means Stump 2 will focus more on it.

* The Correctly classified Sample has Comparatively lower Sample weight
  → which means stump 2 will focus less on it.

* Similarly we can have Many such Stumps

* This is how Adaboost Creates and Uses Stumps.

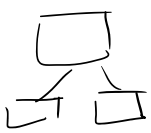
* Let us chk how forest of stumps made by Adaboost does Classification →

(1) Let us consider a test sample

(2) we will pass the test Sample through each stump

(3) Let there be few stumps that Classifies Sample as Yes (Has Heart Disease) and few stumps Classifies Sample as No (Do not have heart disease).

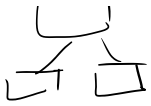
(4) Has Heart Disease



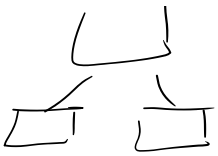
⇒ Amt of Say = 0.47

Has Not Heart Disease

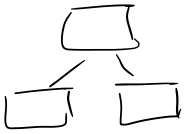
= Amt of Say = 0.41



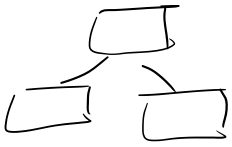
Say = 0.97



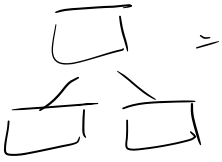
= Amt of Say = 0.82



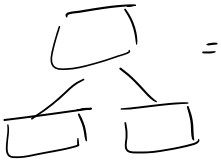
= Amt of Say = 0.32



Total Amt of Say (NO) = 1.23.



= Amt of Say = 0.78



= Amt of Say = 0.63.

Total amt of Say (Yes) = 2.7

Total Amt of Say for Stumps that classifies as Heart Disease (Yes) is greater

∴ The Test Sample is classified as Yes (Has Heart Disease).

Summary of Adaboost

1. Assign Sample Weight (Initial Equal)
2. Create Stump for each feature.
3. Use Gini Index to Identify first stump.
4. For first stump.
   1. Calculate Total Error
   2. Amt of Say.

(4) For first stump.
   (1) Calculate Total Error
   (2) Amt of say.

(5) Update the sample weights for each sample

For Incorrectly classified = updated weight = old weight $* e^{+\text{Amt of say}}$.

For Correctly classified = updated weight = old weight $* e^{-\text{Amt of say}}$.

(6) Now Specify updated weight for each sample & Normalize the weight to generate New Sample weight.

(7) Identify New stump based on New Sample weight & Repeat step 3 to 7

(8) For Classification (Testing)

8.1 ⇒ Run the test sample through all the stumps.

8.2 ⇒ Calculate Amt of say for sample classifying Yes & No

8.3 ⇒ Classify the test sample based on largest sum of Amount of say