XGBoost

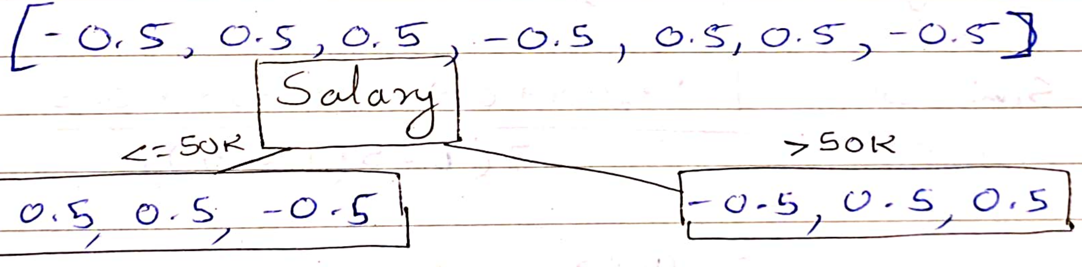
lets consider a loan approval problem where candidates are either approved or rejected based on their salary and credit score

Salary	Credit	Approval	Res
<= 50K	Bad	0	-0.5
<= 50K	Good	1	0.5
<= 50K	Good	1	0.5
> 50K	Bad	0	-0.5
>50K	Good	1	0.5
> 50K	Neutral	1	0.5
<=50K	Neutral	0	-0.5

initial probability = 0.5

Residual's = actual - predicted = $y - \hat{y}$

Our base learners are decision trees, lets see our first stump [trees with depth=1],



we have split on salary with two catogories being <=50K and > 50K, any stump ~~wod~~ would be having only 2 leaves.

Now we calculate similarity score for each leaf and the root.

$$\text{Similarity weight/score} = \frac{\Sigma(\text{Residual})^2}{\Sigma(\text{Prob}(1-\text{Prob})) + \lambda}$$

here is the calculation for leaf <=50K,

$$\text{Simi wt.} = \frac{[-0.5 + 0.5 + 0.5 - 0.5]^2}{0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0.5) + 0.5(1-0}$$

note: initially the probability for all has been initialized as <u>0.5</u> but this will change moving forwar
We also consider λ to be 0 for simplicity.
λ is a hyperparameter.

$$\text{Simi wt} = \frac{0^2}{1} = 0$$

now for leaf >50K,

$$\text{Simi wt} = \frac{[-0.5+0.5+0.5]^2}{0.5(1-0.5)+0.5(1-0.5)+0.5(1-0.5)}$$

$$= 0.33$$

now for the root,

$$\text{Simi wt} = \frac{[-0.5+0.5+0.5-0.5+0.5+0.5-0.5]^2}{7(0.5(1-0.5))}$$

$$= \frac{0.25}{1.75}$$

$$= 0.14$$

now that we have similarity scores for all, we can calculate the Gain for the stump.

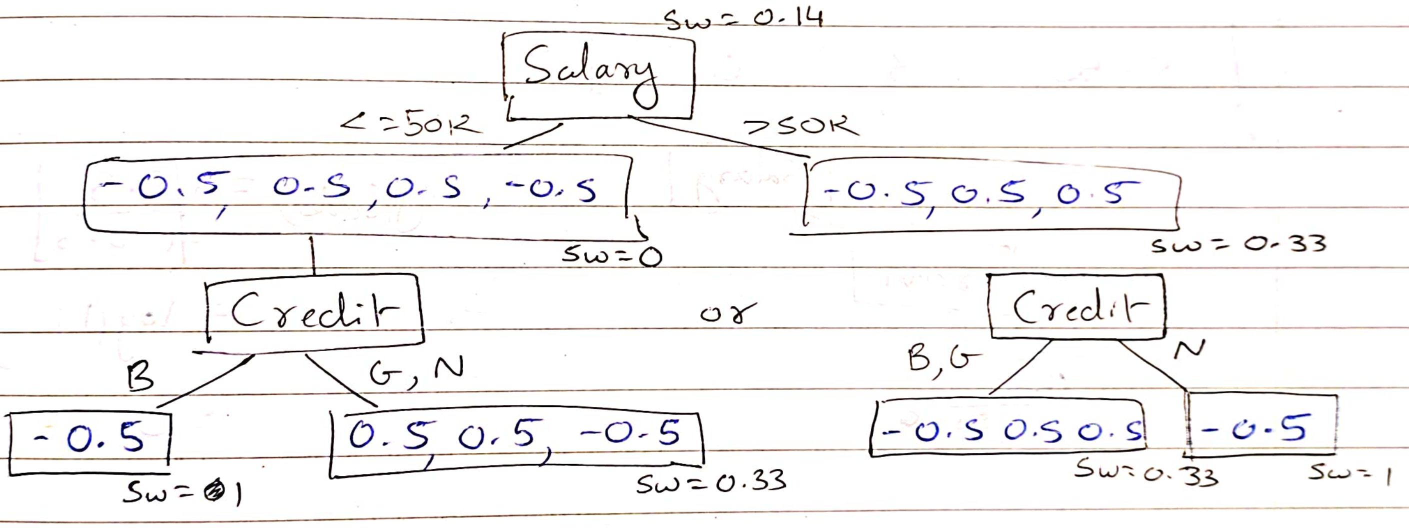
$$\text{Gain} = \text{Similarity}_{\text{left}} + \text{Similarity}_{\text{right}} - \text{Similarity}_{\text{root}}$$

$\therefore$ gain = 0.0 + 0.33 - 0.14

gain = ~~0.21~~ 0.19

the reason we calculate gain is to select the best stump or the stump with the highest gain.
In this problem, we can also split by credit score and if the stump of that has a greater value then we would choose credit score to be our first split.

lets consider that salary was the split with the highest gain. Now we will be splitting with respect to credit.



There are two ways to split using credit and we find gain for both approaches and choose the one with higher gain.

In this case, both variations give the same gain so we can proceed with either of them.

now we need to calculate new probabilities for each entry.

$$\text{new prob} = \sigma\left(\overset{\text{Base model}}{\log(\text{odds})} + lr\,(\cancel{\text{Sim}}\ \text{Similarity weig}\right.$$

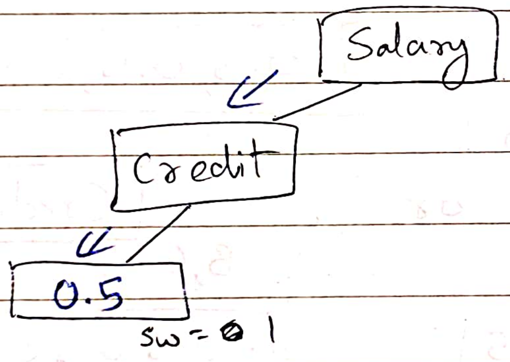
$$\sigma(x) = \frac{1}{1+e^{-x}}$$

lr → learning rate [0 to 1] let = 0.1

$$\log(\text{odds}) = \log\left(\frac{P}{1-P}\right)$$

lets calculate this for the first entry,

[<=50 B 0]



$$\log(\text{odds}) = \log\left[\frac{0.5}{1-0.5}\right] = \log(1) = 0$$

$$\therefore \text{new prob} = \sigma(0 + 0.1(1)) = \sigma(0.1) = \frac{1}{1+e^{-0.1}}$$

$$\boxed{\text{new prob} = 0.6}$$

We do this for each row and calculate the new probability. ~~for~~

our table with new values looks like this,

Salary	Credit	Approval	Res	new Prob	new Res
<= 50K	B	0	-0.5	0.6	-0.4
<= 50K	G	1	0.5	0.4	0.6
<= 50K	G	1	0.5	0.3	0.7
> 50K	B	0	-0.5	0.7	-0.7
> 50K	G	1	0.5	0.2	0.8
> 50K	N	1	0.5	0.1	0.9
<= 50K	N	0	-0.5	0.6	-0.4

we now repeat the same process of creating stumps and calculating new values.

the new probability equation would look like

$$= \sigma(0 + \alpha(T_1) + \alpha(T_2))$$

α = learning rate

this can be generalized as,

$$= \sigma(0 + \alpha(T_1) + \alpha(T_2) \ldots \alpha(T_n))$$

We do this process until we hit our stopping criterion.