~~KNN~~

XGBoost Classifier

→ updated Pr = 0.6

Salary	Credit	Approval	Res	Pr	Res	Base model
<=50K	B	0	-0.5			Pr = 0.5
<=50K	G	1	0.5			
<=50K	G	1	0.5			
>50K	B	0	-0.5			
>50K	G	1	0.5			
>50K	N	1	0.5			
<=50K	N	0	-0.5			

$\sigma(0 + (0.1)(1))$ SW ← 1th record; 0.1 ← learning rate

$\sigma(0.1) = \frac{1}{1+e^{-0.1}} = 0.6$

$Pr = \sigma[0 + \alpha(T_1) + \alpha(T_2) + \alpha(T_3) + \dots \alpha_n(T_n)]$ ✓

eta

α → learning rate

Construct a DT

[-0.5, 0.5, 0.5, 0.5, 0.5, 0.5, -0.5] SW = $\frac{0.25}{7 \times 0.25} = 1/7 = 0.14$

Salary SW = 0.14

<=50K: [-0.5, 0.5, 0.5, -0.5] SW = 0

>50K: [-0.5, 0.5, 0.5] SW = 0.33

$SW = \frac{[-0.5+0.5+0.5+(-0.5)]^2}{\Sigma(Pr(1-Pr)+\lambda)} = 0$

$SW = \frac{0.25}{0.75} = 1/3 = 0.33$

1. Construct tree with Root → gamma
2. Calculate $SW = \frac{\Sigma(\text{Residual})^2}{\Sigma(Pr(1-Pr) + \lambda)}$ → λ hyperparameter

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

3. Calculate gain

$\text{gain} = (SW)_{left} + (SW)_{right} - (SW)_{root}$

$= 0.21$

We compare gain of different features and choose one for which we get most of the gain

Post pruning in XGB classifier = $\frac{Pr(1-Pr)}{\text{cover value}}$ > ~~gain~~ gain → cut the branch

Post pruning

$$Pr = \sigma\left[\log\left(\frac{P}{1-P}\right) + \alpha(SW_1)\} + \alpha(SW_2) + \cdots\right]$$

$$Pr = \sigma\left[\log\left(\frac{0.5}{0.5}\right) + \alpha(1)\right]$$

$$= \sigma\left[0 + 0.1(1)\right] \Rightarrow \frac{1}{e^{0.1}+1} = 0.6$$

XGBoost CART

Imbalanced dataset

Balanced and Imbalanced

Imbalanced dataset may lead to provide a biased output for majority of the cases.

① **Down Sampling**

Take all minority points and sample equal majority points.

Disadvantage:

① Loosing some critical information.

② **Upsampling**

Repeat minority points. Disadvantage:- overfitting

↓

③ Another type → Create artificial / Synthetic datapoints. → extra-polation

↓

another type → Class weight - assign weight in reverse ratio