Performance Measurement of Models

* Accuracy: $\frac{\#\text{ correctly classified points}}{\text{Total \# of pts in } D_{test}}$

Ranges from 0 (Bad) to 1 (Good).

It is easy to understand.

⇒ It should not be used if Data is imbalanced.
⇒ It can't be used to compare 2 models if prob-score is different.

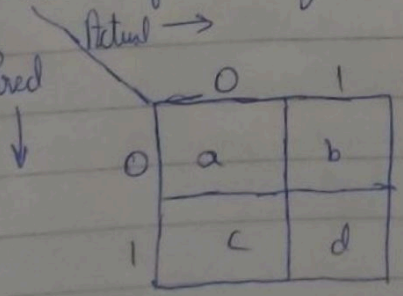
X	y	M_1 (Probability score)	M_2 (Probability score)	$\hat{y}_1$	$\hat{y}_2$
x_1	1	0.9	0.6	1	1
x_2	1	0.8	0.65	1	1
x_3	0	0.1	0.45	0	0
x_4	0	0.15	0.48	0	0

Accuracy M_1 = Accuracy M_2

On the basis of probability scores, $M_1 > M_2$, M_1 model is better than M_2.

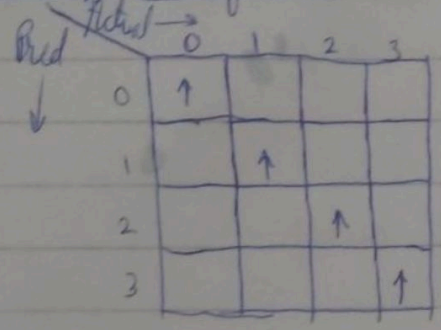
* Confusion Matrix

In case of Binary Classification Task (0, 1)

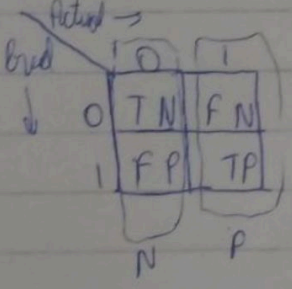


a → no. of pts such that $y = 0$ & $\hat{y} = 0$
b → " " " " " $y = 1$ & $\hat{y} = 0$

Multiclass classification



→ Principal diagonal elements ↑
→ Off diagonal element should be ↓



T P
↳ What is predicted label
↳ Prediction is correct or not

N → No. of −ve pts
P → No. of +ve pts

True Negative Rate, TNR = $\frac{TN}{N}$

False Positive Rate, FPR = $\frac{FP}{N}$

False Negative Rate, FNR = $\frac{FN}{P}$

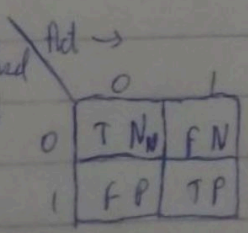
True Positive Rate, TPR = $\frac{TP}{P}$

TNR & TPR should be ↑

FNR & FPR should be ↓

Out of TNR and TPR, what should be Higher?
Ans. It is very domain specific.

* Precision, Recall and F1-Score (They are more focused on +ve class)



Precision = $\frac{TP}{FP+TP}$

(Points predicted to be true out of points that are actually true out of pts that are predicted to be true.)

$$\text{Recall} = \text{TPR} = \frac{TP}{TP+FN}$$ (No. Correctly predicted +ve pts out of total actual +ve pts.)

F1-Score :- Harmonic mean of Precision and Recall.

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Receiver Operating Characteristic Curve (ROC) and AUC
(Mostly used for Binary classification)

X	y	$\hat{y}$
x_1	1	0.95
x_2	1	0.92
x_3	0	0.80
x_4	1	0.76
x_5	1	0.7

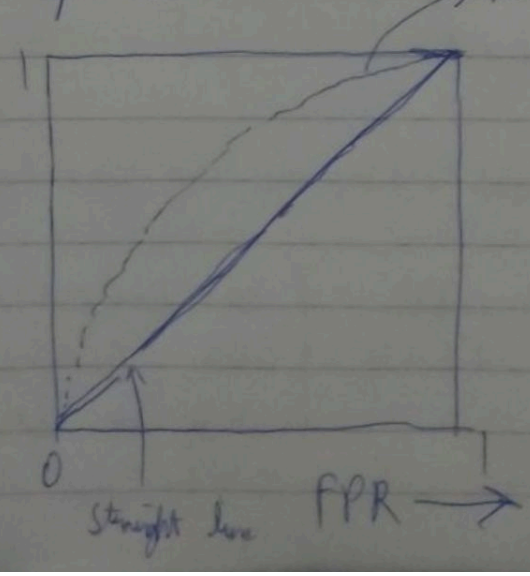
→ First y then $\hat{y}$ are arranged in decreasing order

⇒ Second, thresholding (z) is done with each value of $\hat{y}$.
, i.e,

if $\hat{y} \geq z_1$
return 1
else
return 0

So, after this thresholding, FPR & TPR are calculated for each value of z or $\hat{y}$.

Then they are plotted.



AUC (Area under curve)
= Area under ROC curve

AUC → 0 to 1
↑ ↑
Terrible V. good

~~AUC of a ...~~

Some Properties of AUC:-

⇒ Imbalanced data ⇒ AUC can be high for a dumb model too.

⇒ AUC is not dependent on $\hat{y}$ scores, it only depends on the ordering of $\hat{y}$ scores.

⇒ AUC of a random model is 0.5.

⇒ If AUC of a model is lesser than 0.5, just swap the predicted labels $0 \longleftrightarrow 1$ to get ~~final~~ $AUC_{final} = 1 - AUC_{initial}$

Log-Loss

- One of the best metric for classification problem.

Range :- $[0, \infty)$, lower the better

X	y_i	$\hat{y}_i = P_i$	log loss
x_1	1	0.9	$-\log(0.9) = 0.0457$
x_2	1	0.6	$-\log(0.6) = 0.22$
x_3	0	0.1	$-\log(0.9) = 0.0457$
x_4	0	0.4	$-\log(0.6) = 0.22$

$$\text{log-loss} = -\frac{1}{n}\sum_{i=1}^{n}\left\{(\log(P_i) * y_i) + (1-y_i) * \log(1-P_i)\right\}$$

(y_i: actual values)

↳ It penalizes the small deviations in probs-score

Multi Class Log loss ⇒ $-\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} y_{ij}\log(P_{ij})$

y_{ij}: 1 if $x \in$ class j, 0 otherwise

P_{ij}: Prob that $x \in$ class j

Median Absolute Deviation of errors

R^2 - Coefficient of determination is not robust to outliers.

$e_i \rightarrow y_i, \hat{y}_i$

mean ← median (e_i) = central value of errors

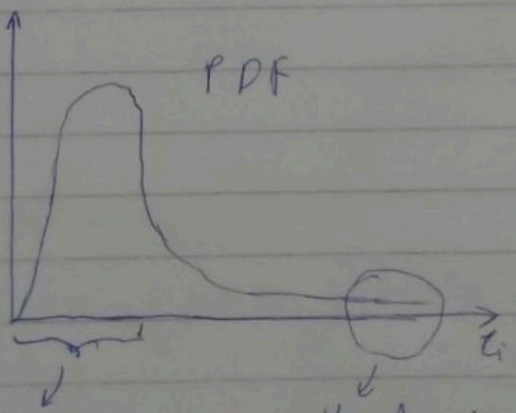
$MAD(e_i)$ = median $(|e_i - \text{median}(e_i)|)$

↙

Standard deviation

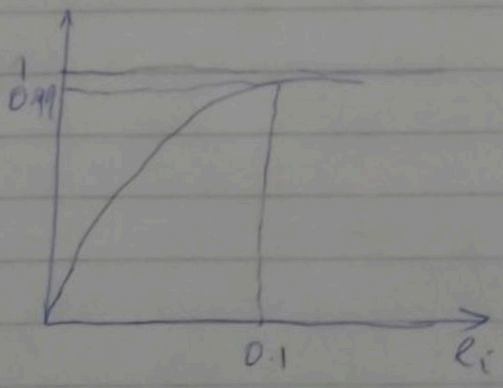
Distribution of errors.

$y_i, \hat{y}_i, e_i$ (errors), we can also create PDF & CDF for e_i's



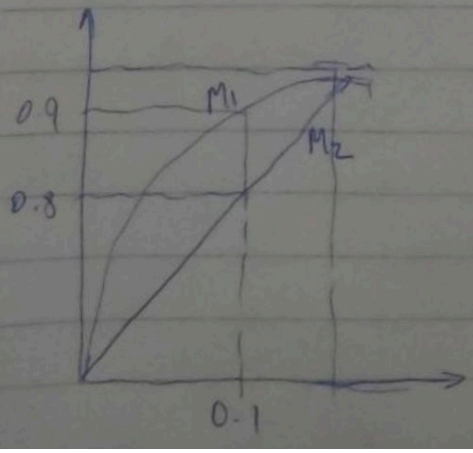
most e_i's are small

Very few e_i's are large



99% of errors are < 0.1

Comparison of Models using CDF :-



For model M_1 :- 90% errors < 0.1

For model M_2 :- 80% errors < 0.1

M_1 is better than M_2.