

METRICS IN CLASSIFICATION

Confusion Matrix - For binomial classification, it is a 2×2 matrix.

Actual Values. Most accurate result = TP and TN

1	0
TP	FP
FN	TN

Type 1 error $\rightarrow FPR = \frac{FP}{FP+TN}$.

Type 2 error $\rightarrow FNR = \frac{FN}{FN+TP}$

AIM OF ANY CLASSIFICATION ML PROBLEM \rightarrow To reduce Type 1 error and Type 2 error

Problem statement

Human being can be pregnant?

Email (Actual)
Spam Not Spam.

Spam	TP	FP
Not Spam	FN	TN

Predicted Value

Actual Value		Testset	
		Cancer	Not Cancer
Female	Actual Value	TP	FP
	Pregnant	Female Pregnant	Male Pregnant
Male	Actual Value	FN	TN
	Not Pregnant	Female Not pregnant	Male Not pregnant

Actual		Cancer		Not Cancer
		TP	FP	Type 1
Female	Actual	FN	TN	Type 2
	Pregnant	Female Pregnant	Male Pregnant	

Type 2 error is more severe than Type 1 error.

1) Accuracy - For balance dataset, we check accuracy. It tells us how many exact cases we predicted. Accuracy = $\frac{TP+TN}{TP+FP+FN+TN}$

For imbalance dataset, we go for Precision, Recall, F Beta score. $\frac{TP}{TP+FP+FN+TN}$.

2) Recall - $\frac{TP}{TP+FN}$, out of all positive values, how many we did we correct predicted as positive. It is also known as true positive rate. Also known as sensitivity. Use Email Spam Case.

3) Precision - $\frac{TP}{TP+FP}$, Out of all predicted positive values, how many are actual positive. Also known as positive prediction value. Use cancer use case.

4) F beta - If we want to consider both Recall and Precision, then use F Beta.

$$F_{\beta} = (1+\beta^2) \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}$$

If $\beta=1$, then it is known as F1 score.

If $\beta=0.5$, then it is known as F0.5 score.

If $\beta=2$, then it is known as F2 score.

$$F_{\beta} = \frac{(2)^{\beta} \times Precision \times Recall}{Precision + Recall}$$

= Harmonic mean

When Type 1 error (FP) and Type 2 (FN) error both are equally important, then choose $\beta=1$.

$$= \frac{2 \times 4}{x+y}$$

When Type 1 error (FP) is more important than Type 2 (FN) error, then reduce β . i.e., $\beta=0.5$.

When Type 2 (FN) is more important than Type 1 (FP) error, increase β . i.e., $\beta=1.5/2$

(Check page 28, for continuity)

ROC and AUC curve

For a probability problem, we always create a threshold value. Above the threshold value we make it one class, and if value is less than threshold value, then another class. By default threshold value = 0.5.

$$TPR = \frac{TP}{TP+FN}$$

Take some threshold value $\rightarrow 0, 0.2, 0.4, 0.6, 0.8, 1$.

Sample.

$$FPR = \frac{FP}{FP+TN}$$

y (Actual)	\hat{y} (Predicted)	$\hat{y}(0)$	$\hat{y}(0.2)$	$\hat{y}(0.4)$
1	0.8	1 (TP)	1 (TP)	1 (TP)
0	0.96	1 (FP)	1 (FP)	1 (FP)
1	0.4	1 (TP)	1 (TP)	0 (TN)
1	0.3	1 (TP)	1 (TP)	0 (TN)
0	0.2	1 (FP)	0 (FN)	0 (FP) (FN)
1	0.7	1 (TP)	1 (TP)	1 (TP)

$$TPR = \frac{1}{1+0} = 1$$

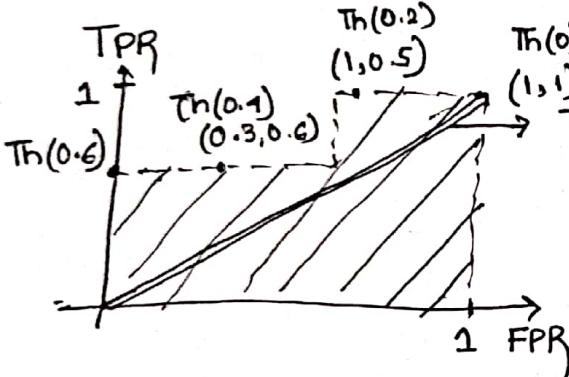
$$FPR = \frac{2}{2+0} = 1$$

$$TPR = \frac{5}{5+1} = 1$$

$$FPR = \frac{1}{1+1} = 0.5$$

$$TPR = \frac{2}{2+1} = \frac{2}{3} = 0.6$$

$$FPR = \frac{1}{1+2} = \frac{1}{3} = 0.3$$



- More the area under the curve, better the model is. This is known as AUC (Area under Curve). The area should be increase than this area (half area) or else it is a dump model (no use model).

- ROC is the Receiver Operating characteristics curve. It is created by plotting the true positive rate against false negative rate at various threshold setting.

CONFUSION MATRIX	2 X 2 Matrix		
Type 1 error	FPR (False Positive Rate) = $\frac{FP}{FP+TN}$		
Type 2 error	FNR, False Negative Rate = $\frac{FN}{FN+TP}$	Type 2 is more dangerous than Type 1. Cancer vs case	
Accuracy.	Balance dataset	$\frac{TP+TN}{TP+FP+FN+TN}$	
Recall	Imbalance dataset True positive rate.	Sensitivity Email spam vs user case	Out of all actual positive, how many we correctly predicted. $\frac{TP}{TP+FN}$
Precision	Imbalance dataset Positive prediction value	Cancer User case	Out of all predicted positive value, how many are actual positive. $\frac{TP}{TP+FP}$
F-Beta.	Consider both Recall and precision.	$F_{\beta} = \frac{(1+\beta^2)}{(1+\beta^2)} \frac{Precision \times Recall}{\beta^2 \times (Recall+Precision)}$	
		If Type 1 = Type 2 important choose $\beta=1$. F_1 score = $\frac{(2)Precision \times Recall}{Precision+Recall}$	
		If Type 1 (important) $>$ Type 2. choose $\beta=0.5$. $F_0.5$ score.	
		If Type 1 $<$ Type 2 (important) choose $\beta=2$. F_2 score.	
Receiver Operating characteristic Area under curve (AUC)	ROC is plot between TPR and FPR Higher the area under the curve better is the model.		