Last class (Aug 19)

1) Spam vs Non-Spam : Business Case
2) Issue with Accuracy
3) Confusion Matrix & Code
4) Precision & Code
5) Recall & Code
6) F1 score & Code

Today's class

1) Recap — Quizzes
2) Sensitivity and Specificity
3) ROC curve
4) AUC under ROC curve
5) Precision Recall Curve
6) Handling Imbalance Data
   - → class weights
   - → Oversampling of minority
   - → Undersampling of majority
   - → SMOTE

Movie reco

→ FP → A bad movie recommended

→ FN → A good movie not recommended

high precision

Malware detection

FP: good software predicted as Malware

→ FN: malware undetected (detected as good software)

TP, TN, FP, FN	equally imp

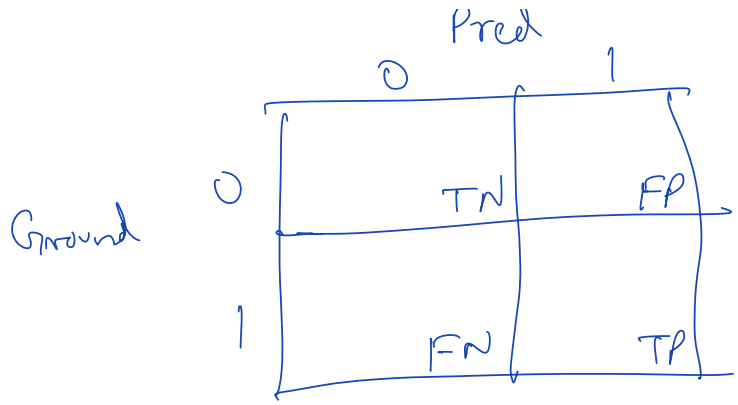
$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \text{Recall} = \text{True Positive Rate}$$

$$\text{Specificity} = \frac{TN}{TN + FP} = 1 - \frac{FP}{FP + TN}$$

$$= 1 - \text{False Positive Rate}$$



Balanced Data

500 → class 0

500 → class 1

TP, FP, FN not sufficient

Bank Fraud Detection → class 0 class 1

1000 1000

Cancer Detection → class 0 class 1

600 800 ✓ ✓

Prec → $\frac{TP}{TP+FP}$

Recall → $\frac{TP}{TP+FN}$

$Acc = \frac{TP+TN}{FP+FN+TP+TN}$

Req → [80% Precision, 90% Recall]

Logistic Regression

$\sigma(z) = p\left(y^{(i)}=1 \mid x^{(i)}\right) = p$

$= 0 \text{ — } 1$

class 0 or class 1 →

$p \geq \boxed{0.5}$ → class 1

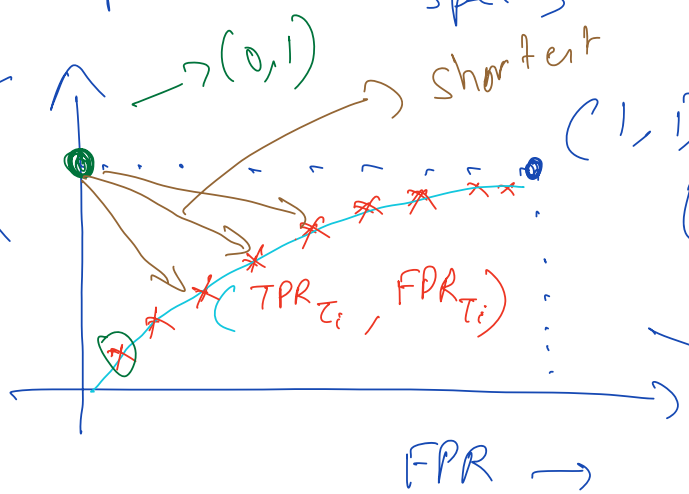
$p < 0.5$ → class 0

$T_1, T_2, T_3, \ldots T_N$

Prec 1
Recall 1
Spec 1

Prec 3
Recall 3
Spec 3

ROC Curve



Max value of TPR = 1
FPR = 1
TPR = 1
FPR = 1

ROC Curve

Ideal Model: $TPR = 1 = \frac{TP}{TP + FN} \to 0$

$FPR = 0 \Rightarrow \frac{FP}{FP + TN} = 0$

$\Rightarrow 0$

Model M_1, Model M_2

Logistic Regression

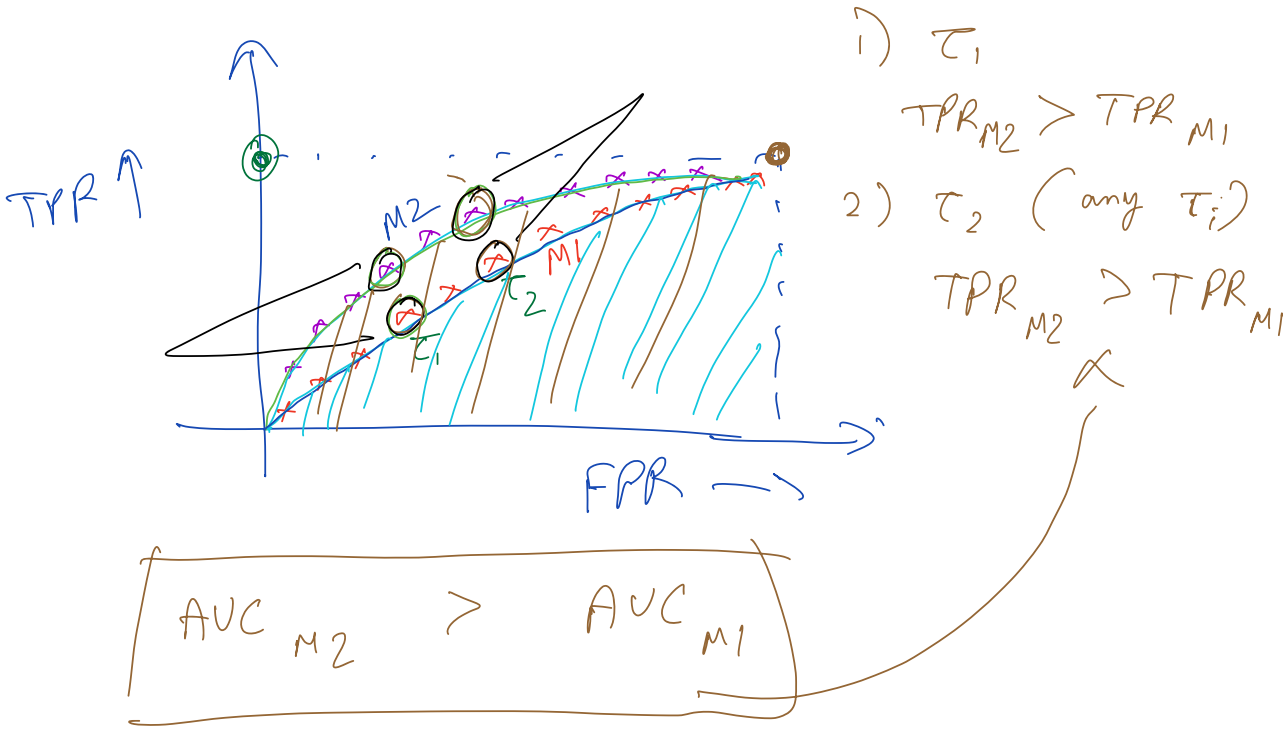
Neural network

$\tau = 0.5$

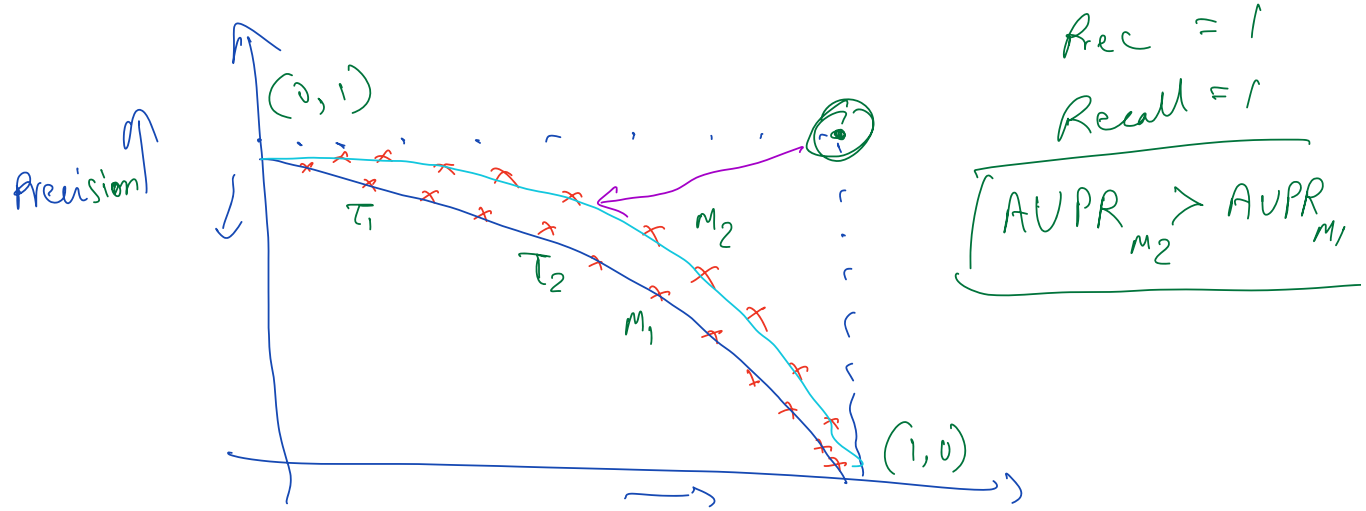
$Acc_1 = 0.8$

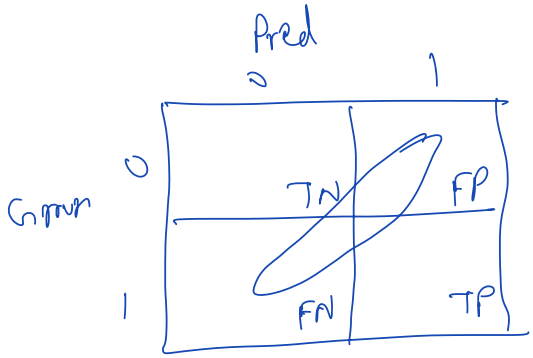
$Acc_1 = 0.8$

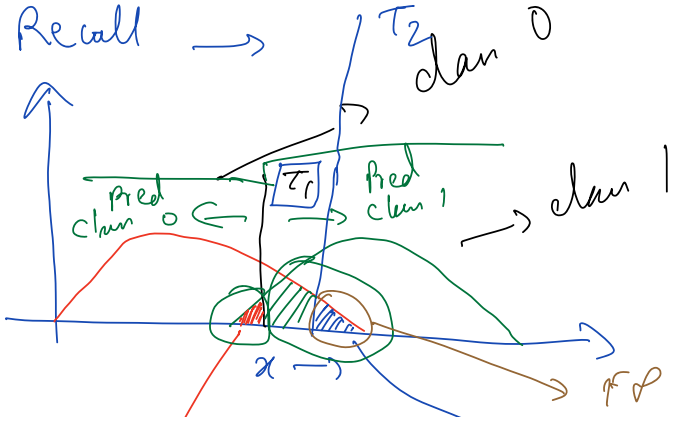
Balanced Dataset



Imbalanced Dataset







τ_1 : $FN << FP$

τ_2 : $FP << FN$

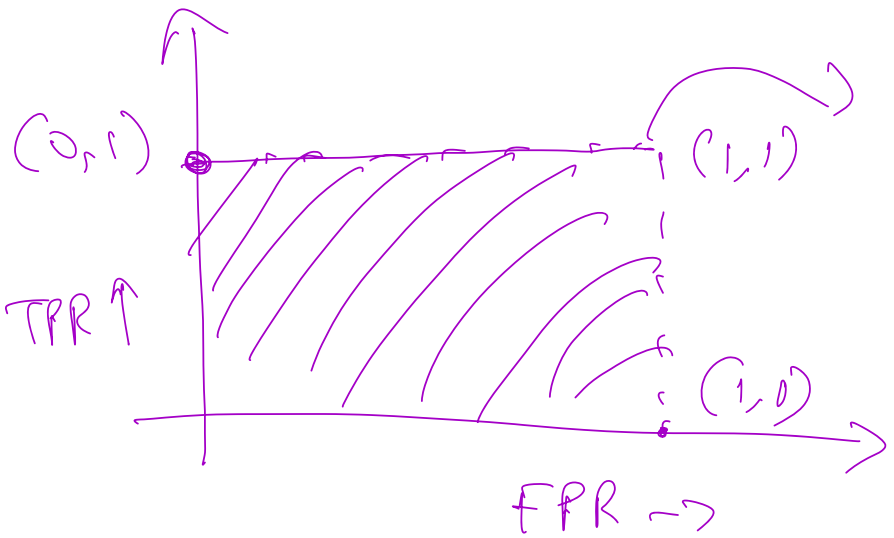
GD : class 1 (+ve)

Pred : class 0 (-ve) — FN

GD : class 0

Pred : class 1 — FP

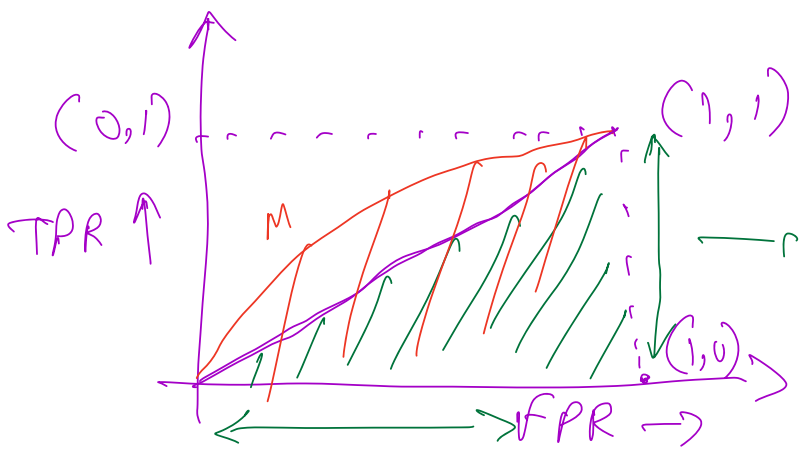
Ideal Model



50% +ve, 50% -ve | Random Model

class 0 : 50% prob

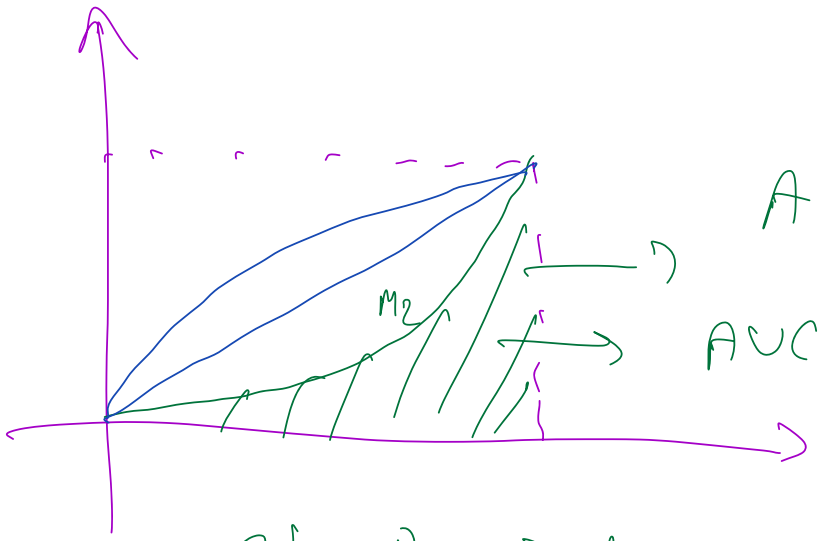
class 1 : 50% prob



$AUC = \frac{1}{2} \times 1 \times 1$

$= 0.5$

$AUC_M > 0.5$



$AUC_{M2} < 0.5$

$AUC_{M2} = 0.4$

$1 - AUC_{M2} = 0.6$

class 0 → 1

class 1 → 0

$X = [x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)}, x^{(5)}, x^{(6)}]$

$Y = [1, 1, 0, 1, 0, 0]$ → 3 +ve, → 3 −ve

$P = [0.65, 0.94, 0.3, 0.92, 0.7, 0.2]$

X	Y	P
$x^{(1)}$	1	0.65
$x^{(2)}$	1	0.94
$x^{(3)}$	0	0.3
$x^{(4)}$	1	0.92
$x^{(5)}$	0	0.7
$x^{(6)}$	0	0.2

Sort them descending order

X	Y	$\hat{y}_{\tau_1}$	$\hat{y}_{\tau_5}$	P
$x^{(2)}$	→ 1	1	1	$0.94 = \tau_1$
$x^{(4)}$	→ 1	0	1	$0.92 = \tau_2$
$x^{(5)}$	→ 0	0	1	$0.7 = \tau_3$
$x^{(1)}$	→ 1	0	1	$0.65 = \tau_4$
$x^{(3)}$	→ 0	0	1	$0.3 = \tau_5$
$x^{(6)}$	→ 0	0	0	$0.2 = \tau_6$

sorted

τ_1 : pred +ve = 1
pred -ve = 5

TP	FP	TN	FN
1	0	3	2

$TPR = \frac{1}{1+2} = 0.33$

$FPR = \frac{0}{0+3} = 0$

→ (0, 0.33)

τ_3 : pred +ve = 3
pred -ve = 3

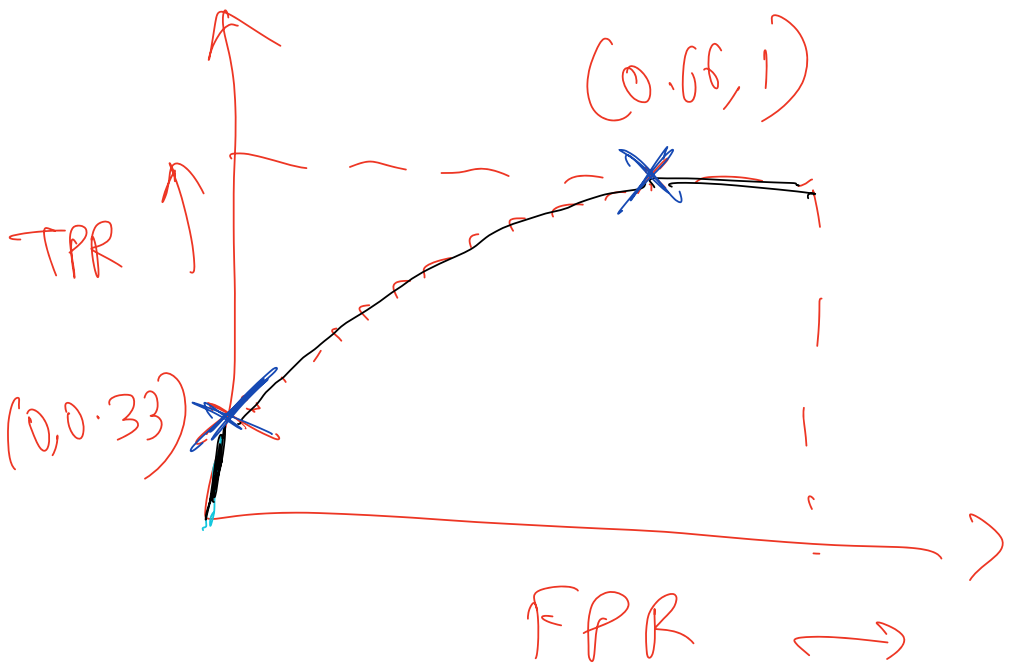
τ_5 : pred +ve = 5
pred -ve = 1

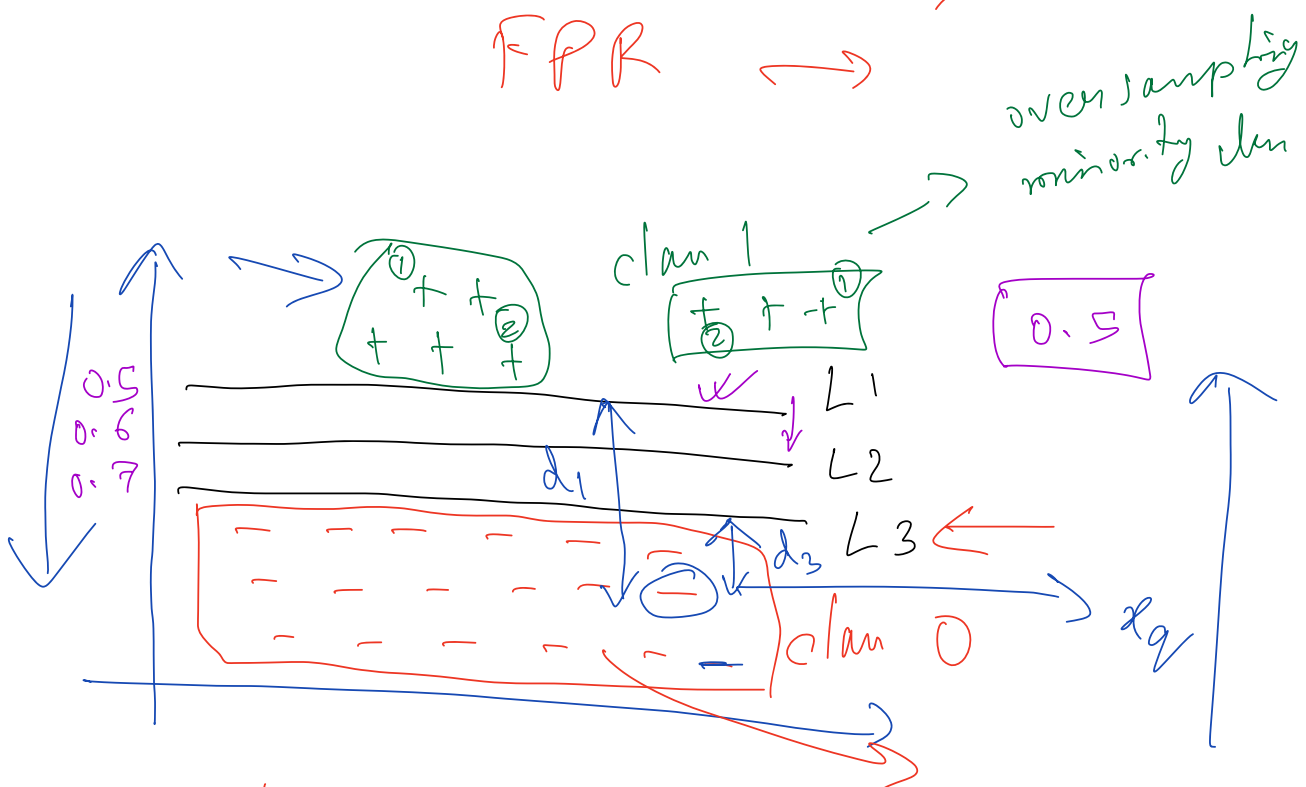
TP	FP	TN	FN
3	2	1	0

$TPR = \frac{3}{3+0} = 1$

$FPR = \frac{2}{2+1} = 0.66$

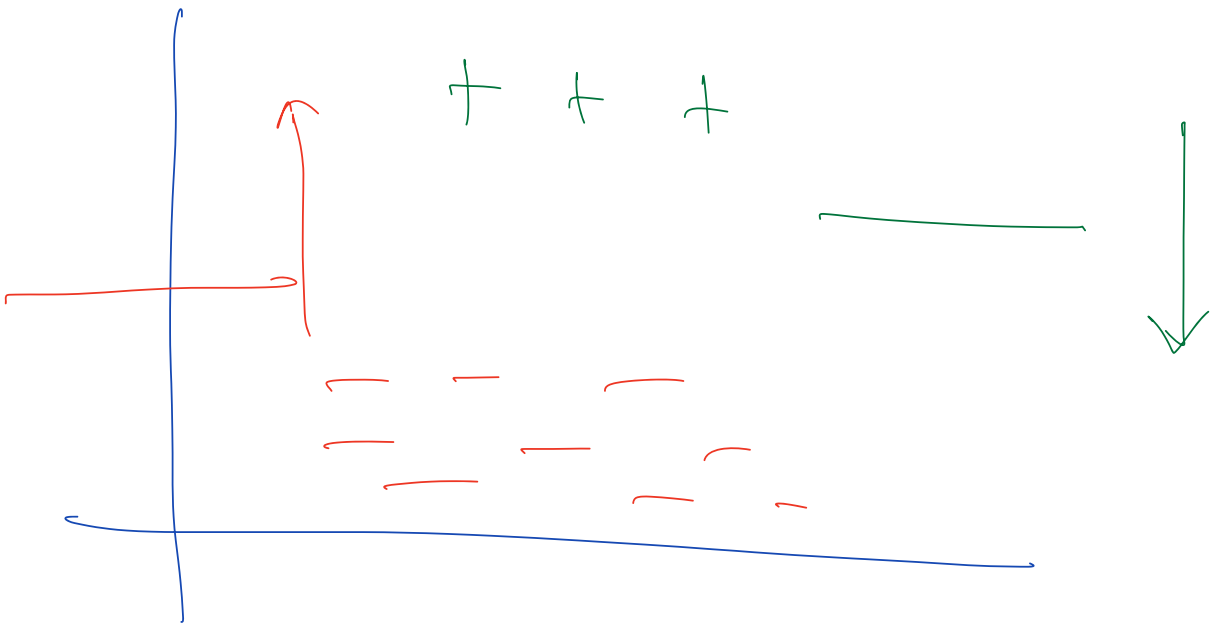
(0.66, 1)





$N_0 >> N_1$

$\rightarrow d_1 > d_3 \qquad c_1 > c_3$



$N_0 = 1000$

$N_1 = 100$

$L(\ ) \rightarrow$ −ve sample $\rightarrow \frac{1}{10}$ of

$L(\ ) \rightarrow$ +ve sample

$w = \frac{1}{10}$ $L_{+,1}$, $L_{+,2}$, $L_{+,100}$ $\rightarrow$ $\leftarrow$ $L_{-,1}$, $L_{-,2}$, ⋮, $L_{-,10}$ $w = 1$

$$w_c = \frac{k}{N_c}$$

$N_0 = 1000$

$N_1 = 100$

$w_0 = \frac{k}{1000}$

$w_1 = \frac{k}{100}$

$w_0 + w_1 = 1$

$k\left(\frac{1}{1000} + \frac{1}{100}\right) = 1$

$\Rightarrow k\left(\frac{1100}{1000 \times 100}\right) = 1$

$N_0 = 100$

$N_1 = 1000$

$N_2 = 5000$

$w_0 = \frac{k}{N_0}$, $w_1 = \frac{k}{N_1}$, $w_2 = \frac{k}{N_2}$

$w_0 + w_1 + w_2 = 1$

$$\text{log-loss } (L) = -\sum_{i=1}^{m}\left[y^{(i)} \log\left(\hat{y}^{(i)}\right) + \left(1-y^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)\right]$$

Weighted log-loss

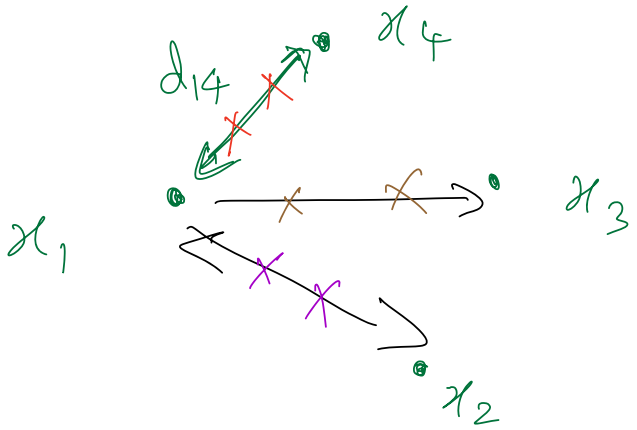
$$= -\sum_{i=1}^{m}\left[w_1\, y^{(i)} \log\left(\hat{y}^{(i)}\right) + w_0\left(1-y^{(i)}\right)\log\left(1-\hat{y}^{(i)}\right)\right]$$

SMOTE

$\varepsilon = (0, 1)$

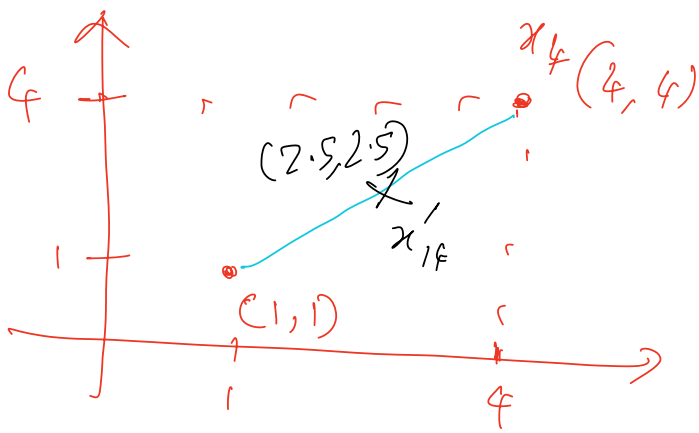
$0.4, 0.7$

$x'_{14} = x_1 + \varepsilon \cdot d_{14}$



$x_1 = (1, 1)$

$x_4 = (4, 4)$



$\varepsilon = 0.5$

$x'_{14} = x_1 + \varepsilon \, d_{14}$

$= (1, 1) + 0.5 \times (4-1, 4-1)$

$= (1, 1) + 0.5 \times (3, 3)$

$= \left(1 + \frac{3}{2}, \quad 1 + \frac{3}{2}\right)$

$= (2.5, \ 2.5)$

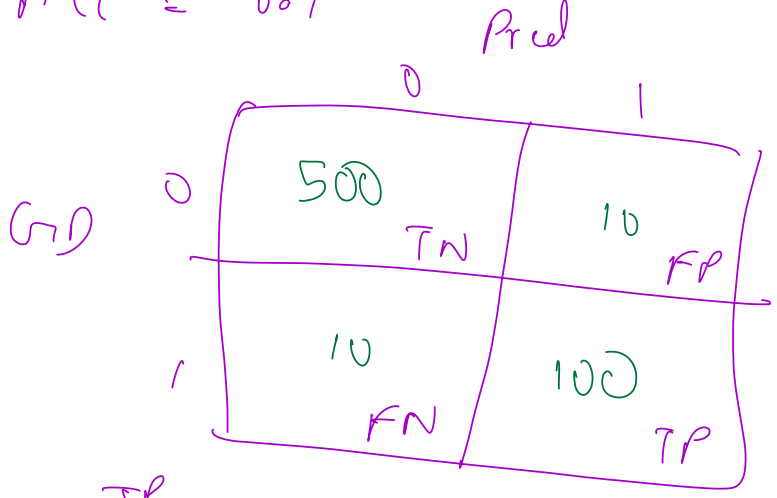
360 +ve 40 −ve

P = 1
R = 0.9

dumb model → TP TN FP FN

F1 = 0.947

Acc = 0.9



$Rec = \frac{TP}{TP + FN} \to 1 = 1$

$Prec = \frac{TP}{TP + FP} \to 1$

$$\text{Prec} = \frac{100}{100+10} = \frac{10}{11} = 0.91$$

$$\text{Recall} = 0.91$$

$$\text{Acc} = \frac{600}{620} = \sim 0.96$$

Balanced data is a sub-set of Imbalanced data

If Prec & Recall is high,

$\Rightarrow$ Acc should be high.