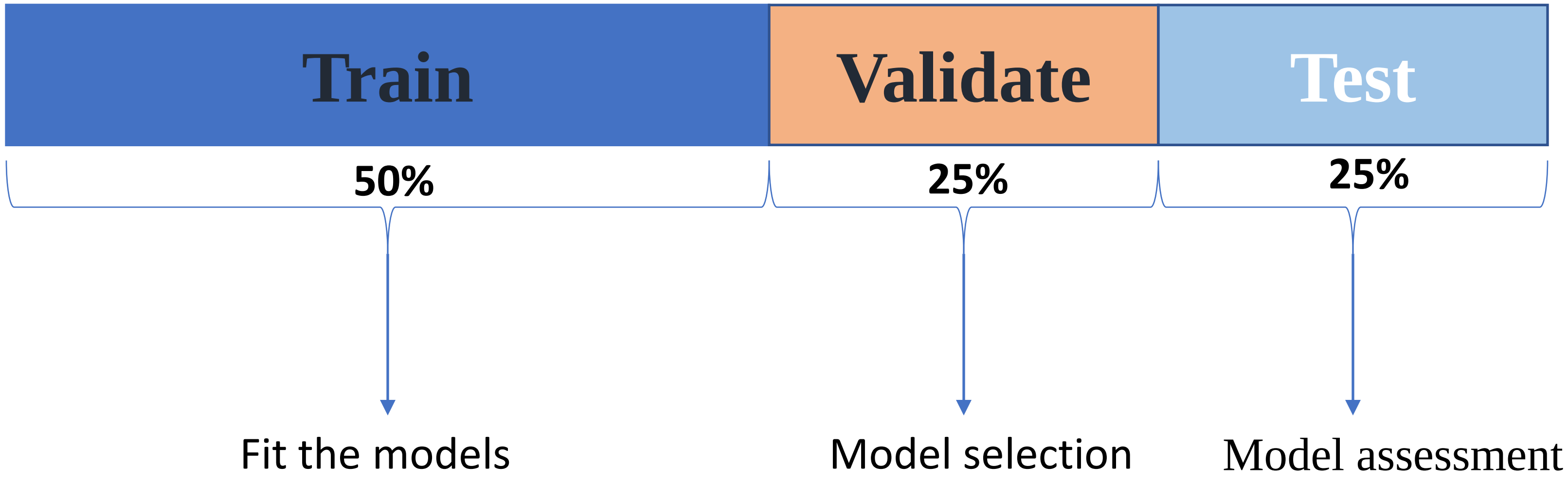
# Module II: Machine Learning: Foundations and Algorithms

Nirav P Bhatt
Department of Biotechnology
Robert Bosch Centre for Data Science and Artificial Intelligence
Indian Institute of Technology Madras, Chennai – 600036, India

# Model Selection and Assessment

- Model selection is important in multiple linear and nonlinear models
- Data-rich situation: Randomly divide the data in three parts

**Ideal Scenario: Data-rich situation**




***Practice: Limited Amount of Data***

*Best Model in Practice? Need a Criterion*

# Model Selection and Assessment

- Development of model
  - **Building a model**(s): Verifying all the assumptions
  - Validation of model: Predictive ability of the model
  - Testing the model on new data
- Two goals: Validation of model
  - **Model selection:** Comparing the performance of several models to find the best one
  - **Model assessment:** Assessing the predictive ability of the chosen final model on new data

# Irreducible and Reducible Errors

Mean Square Error between the actual and predicted $y$ *using* the fit $\hat{f}(x, \hat{p})$

$$E[(y - \hat{y})^2] = [f(x,p) - \hat{f}(x,\hat{p})]^2 + Var(\epsilon)$$

Irreducible Error $Var(\epsilon)$

Reducible Error $[f(x,p) - \hat{f}(x,\hat{p})]^2$

# Bias-Variance Trade-off

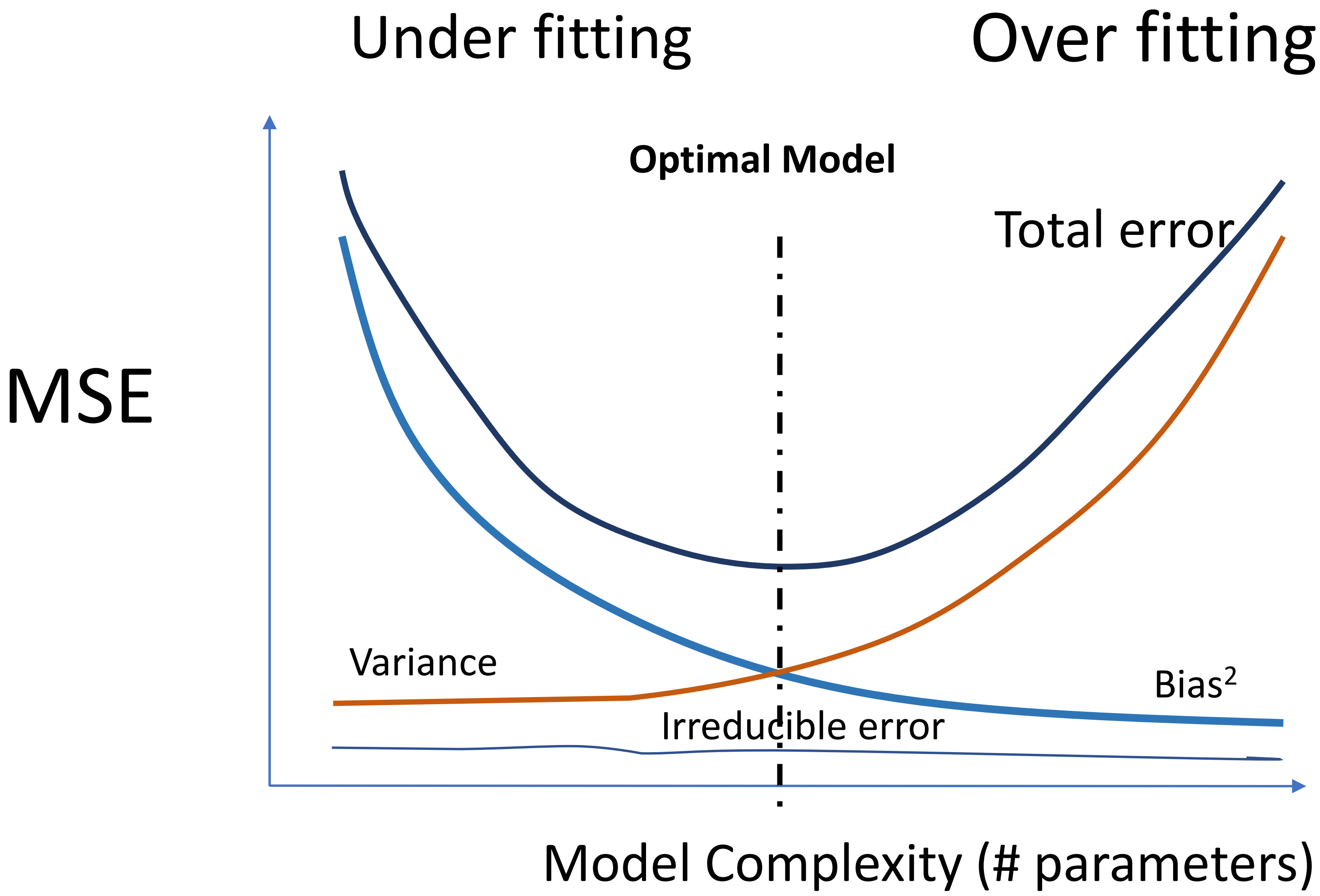
kNN MSE

$$E[(\hat{y}_{x_o} - y)^2] = Var(\epsilon) + \underbrace{\frac{1}{K}\sigma^2}_{\text{Variance}} + \underbrace{(f(x_o) - \frac{1}{K}\sum_{i \in \mathcal{A}} f(x_i))^2}_{\text{Bias}}$$

Linear Regression MSE

$$E[(\hat{y} - y)^2] = \sigma^2 + \underbrace{(\mathbf{x}_p^T Var[\hat{\boldsymbol{\beta}}_p]\mathbf{x}_p)}_{\text{Variance}} + \underbrace{(\mathbf{x}_p^T \mathbf{A}\boldsymbol{\beta}_r - \mathbf{x}_r \beta_r)^2}_{\text{Bias}}$$

# Bias-Variance Trade-off



# Model Validation: Methods

## Model Validation

selections + hyper-parameters

**Analytical Methods**
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Mallows Cp

**Resampling Methods**
- Validation set approach
- K-fold  cross validation
- Bootstrap

Model Selection & Assessment

**Subset selection**
- Best subsets
- Backward elimination
- Forward Selection
- Stepwise regression

Feature Engineering

# Classification Models

$\{x_n^1, x_n^2, \dots, x_n^p\}$

p features

- Data: $(x_1,y_1), (x_2,y_2),\dots,(x_n,y_n)$
  - Binary class problems $\longrightarrow \{0,1\}$
  - Multi-class problems $\longrightarrow \{0, 1, \dots k-1\}$ k classes
- Underlying true distribution P(X, y)
- How well the underlying distribution learnt by a Classifier?
- Questions
  - How do we estimate the true performance of a classifier?
  - How good are the parameter estimates in the classifier?

# Evaluation Metrics: Binary Classification

| T | + | + | + | + | - | - | - | - | - | - |
|---|---|---|---|---|---|---|---|---|---|---|
| P | + | - | + | - | + | - | - | - | - | - |
| Outcome | TP | FN | TP | FN | FP | TN | TN | TN | TN | TN |

TP: True Positive (Positive sample classified as positive class)
FN: False Negative (Positive sample classified as a negative class)
FP: False Positive (Negative sample classified as positive class)
TN: True Positive (Negative sample classified as negative class)

# Evaluation Metrics: Binary Classification

**Confusion Matrix (Contingency table)**

| | | **Predicted** | |
|---|---|---|---|
| | | **+** | **–** |
| **True** | **+** | True Positive (TP) | False Negative (FN) |
| | **–** | False Positive (FP) | True Negative (TN) |

$$\text{Accuracy} = \frac{TP+TN}{P+N}$$

Misclassification rate=1- Accuracy

Note: In the literature, the confusion matrix can be the transpose of the present (CM), i.e. true classes is presented by the column & the predict one in the row.

# Evaluation Metrics: Binary Classification

| | + | – |
|---|---|---|
| **+** | 15 | 5 |
| **–** | 30 | 950 |

| | + | – |
|---|---|---|
| **+** | 18 | 2 |
| **–** | 20 | 960 |

Compute Accuracy

Accuracy = $\frac{15+950}{20+980} = 0.965$

Accuracy = $\frac{18+960}{20+980} = 0.978$

Accuracies for both classifier are quite identical.

For Highly imbalance data, Accuracies is not a good measure.

# Evaluation Metrics: Binary Classification

**Confusion Matrix (Contingency table)**

| | | Predicted + | Predicted - |
|---|---|---|---|
| **True** | + | True Positive (TP) | False Negative (FN) |
| | - | False Positive (FP) | True Negative (TN) |

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

$$\text{Precision} = \frac{\begin{gathered} Positive\ samples\ are\ classified \\ as\ positive\ by\ a\ classifier \end{gathered}}{\begin{gathered} Total\ positive\ class\ samples \\ predicted\ by\ a\ classifier \end{gathered}}$$

*The fraction of true positive classes corrected predicted by a classifier in all the positive classes predicted by the classifier*

# Evaluation Metrics: Binary Classification

**<u>Confusion Matrix (Contingency table)</u>**

| | | **Predicted** | |
|---|---|---|---|
| | | **+** | **–** |
| **True** | **+** | True Positive (TP) | False Negative (FN) |
| | **–** | False Positive (FP) | True Negative (TN) |

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

The fraction of true positives predicted by a Classifier wrt to the true positive.

# Evaluation Metrics: Binary Classification

| | + | – |
|---|---|---|
| + | 15 | 5 |
| – | 30 | 950 |

| | + | – |
|---|---|---|
| + | 18 | 2 |
| – | 10 | 970 |

Precision

$$\frac{15}{15+30} = 0.67$$

$$\frac{18}{18+10} = 0.65$$

Recall

$$\frac{15}{15+5} = 0.75$$

$$\frac{18}{18+2} = 0.9$$

# Evaluation Matrix: Binary Classification

**Confusion Matrix (Contingency table)**

| | | **Predicted** + | **Predicted** - |
|---|---|---|---|
| **True** | + | True Positive (TP) | False Negative (FN) |
| **True** | - | False Positive (FP) | True Negative (TN) |

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

Recall, sensitivity = $\frac{TP}{TP+FN}$

specificity : How well a diagnostic test can detect TN?

sensitivity : How well a diagnostic test can detect TP?

# Evaluation Metrics: Binary Classification

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

Recall, sensitivity = $\frac{TP}{TP+FN}$

RT-PCR vs Cancer (Mammogram) Test

RT-PCR

→ covid +ve

→ covid −ve

specificity: The RT-PCR's ability to correctly reject covid-ve person

Benign or Cancerous tumor

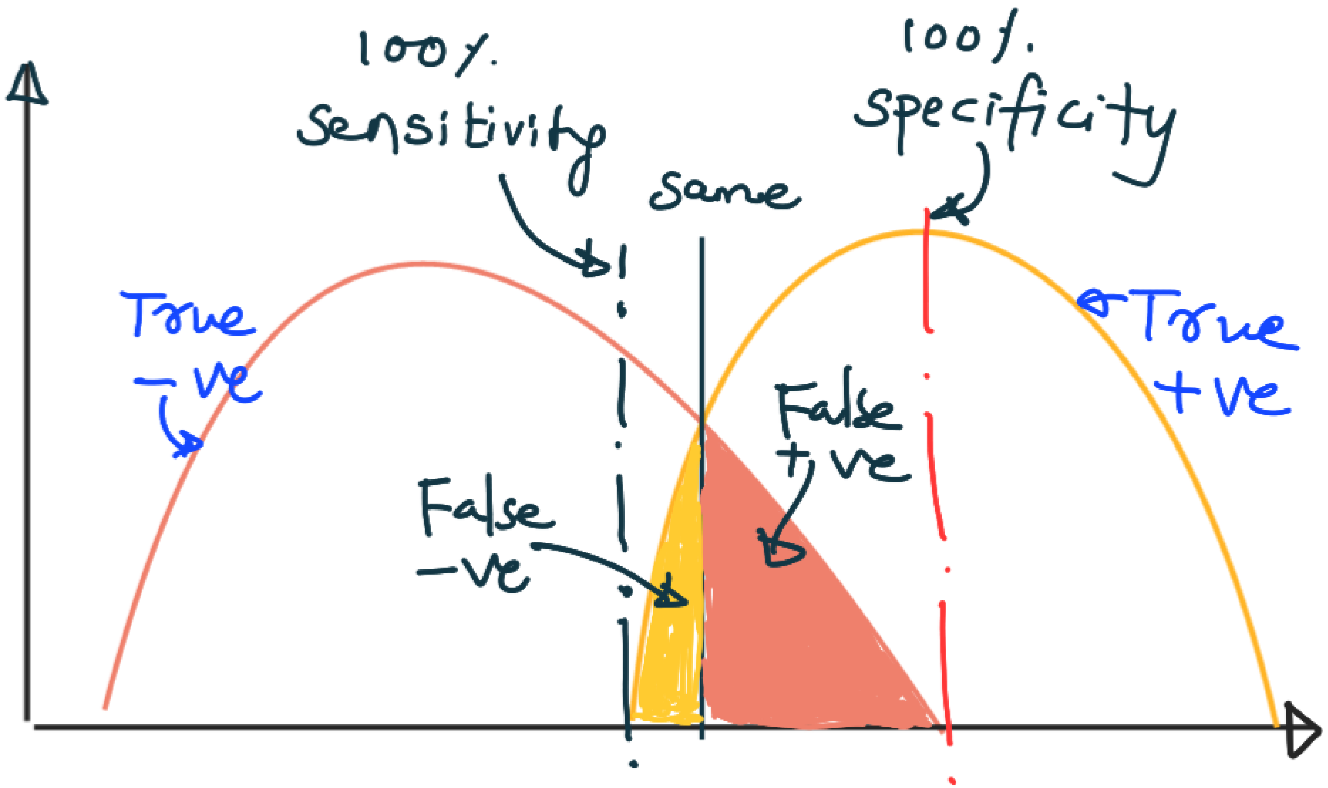
Sensitivity: Mammogram based test's ability to correctly detect cancerous tumor.

# Evaluation Metrics: Binary Classification

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

Recall, sensitivity = $\frac{TP}{TP+FN}$

RT-PCR vs  Cancer (Mammogram) Test



# Evaluation Metrics: Binary Classification

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

Recall, sensitivity = $\frac{TP}{TP+FN}$

RT-PCR vs Cancer (Mammogram) Test

Blank slide

# Evaluation Metrics: Binary Classification

**Receiver Operating Characteristic (ROC) Curve**

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

False Positive Rate = $\frac{FP}{FP+TN}$

**(ROC) Curve:**
**A graph between FPR vs TPR**

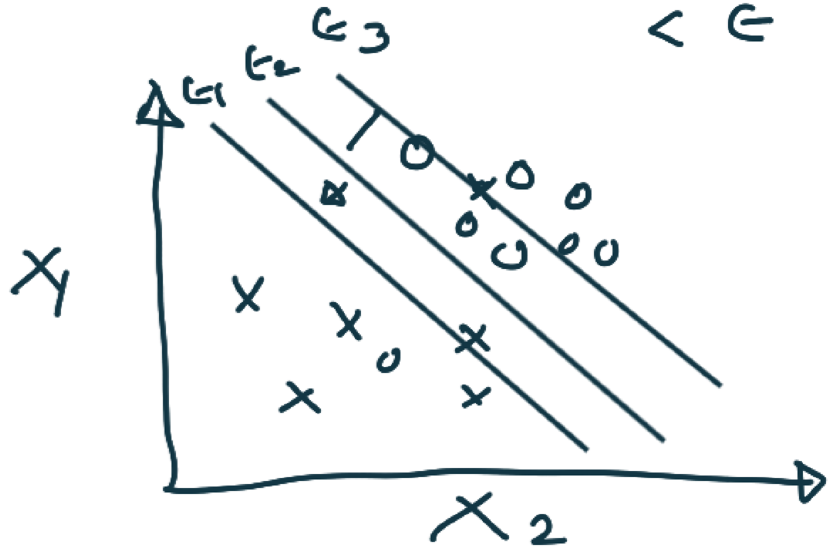
Classifiers

* Thresholds

↳ Probabilities

$P(y/X) > \epsilon$ , +

$< \epsilon$ , −

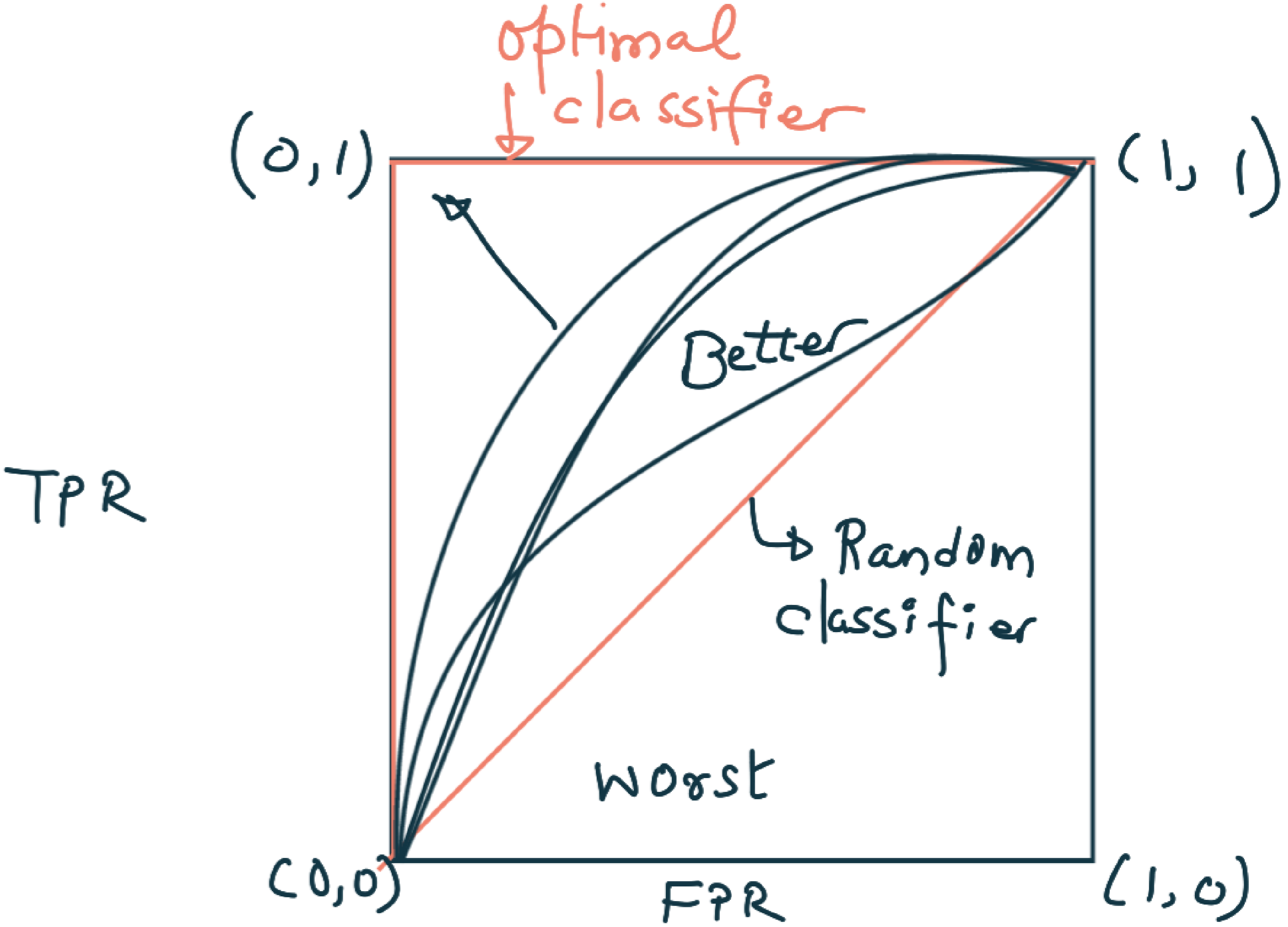


# Evaluation Matrix: Binary Classification

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

False Positive Rate = $\frac{FP}{FP+TN}$

**Receiver Operating Characteristic (ROC) Curve**



# Evaluation Metrics: Binary Classification

**Receiver Operating Characteristic (ROC) Curve**

$$\text{True Positive Rate} = \frac{TP}{TP+FN} \times 100$$

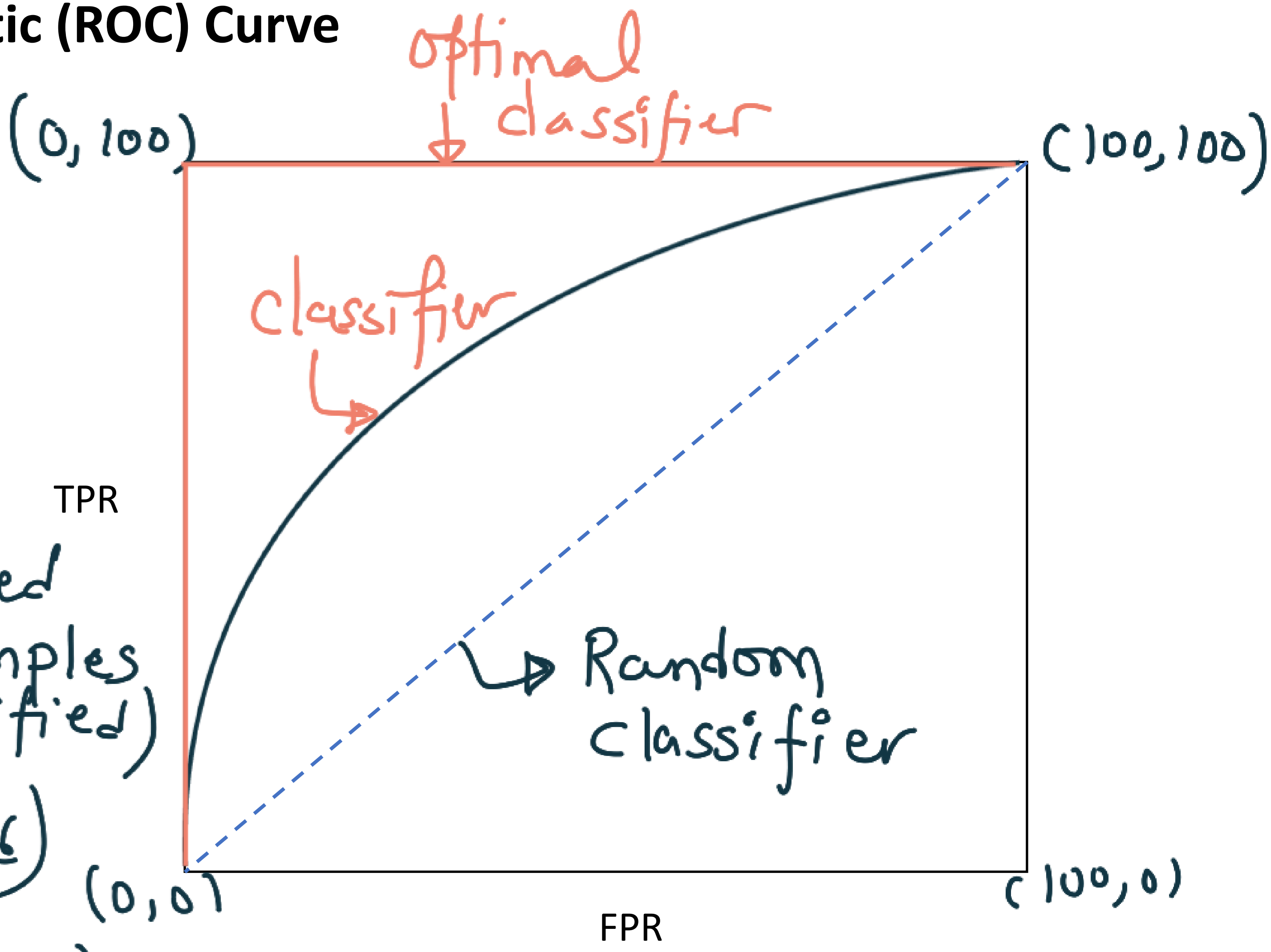
$$\text{False Positive Rate} = \frac{FP}{FP+TN} \times 100$$

(0,0) → −ve samples correctly classified

(100,100) → (+ve/+ve, −ve samples misclassified)

(100,0) → (+ve & −ve misclassified)

(0,100) → (+ve & −ve classified correctly)



# Evaluation Metrics: Binary Classification

## Receiver Operating Characteristic (ROC) Curve

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

False Positive Rate = $\frac{FP}{FP+TN}$

Example

| Points | Probability | Threshold | TPR | FPR |
|---|---|---|---|---|
| + | 0.9 | 0.9 | 1/4 | 0 |
| + | 0.85 | 0.8 | 2/4 | 0 |
| + | 0.6 | 0.5 | 3/4 | 0 |
| - | 0.49 | 0.45 | 3/4 | 1/6 |
| - | 0.4 | 0.39 | 3/4 | 2/6 |
| - | 0.35 | 0.35 | 3/4 | 3/6 |
| - | 0.34 | 0.33 | 3/4 | 4/6 |
| + | 0.32 | 0.3 | 1 | 4/6 |
| - | 0.2 | 0.2 | 1 | 5/6 |
| - | 0.1 | 0.1 | 1 | 1 |

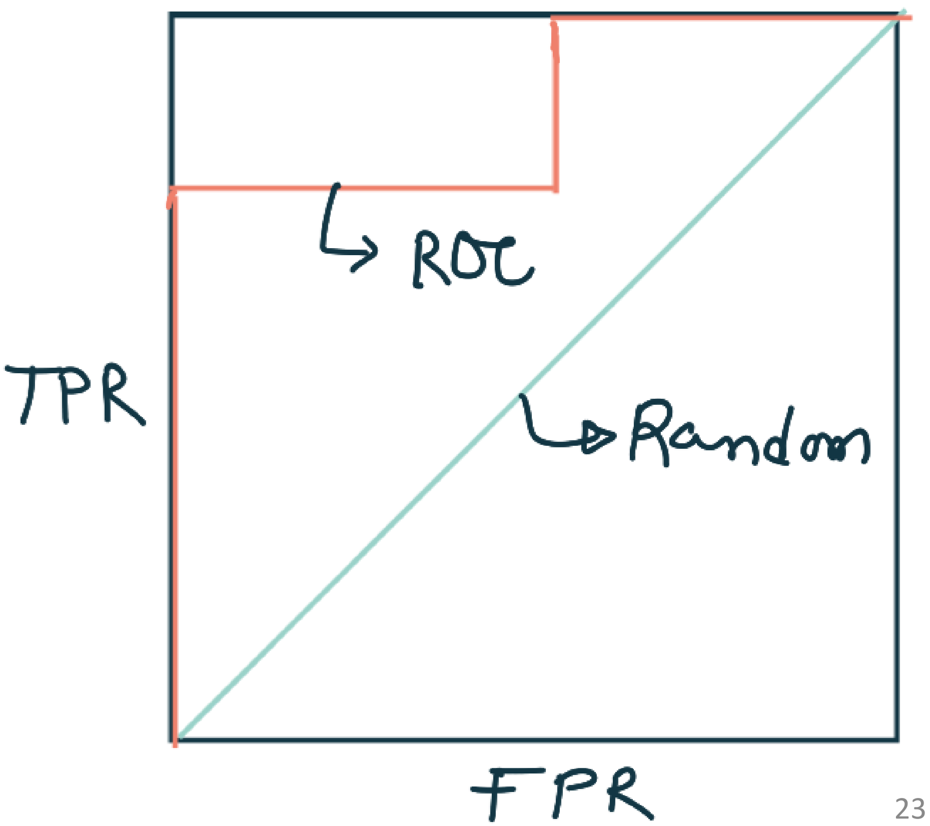
# Evaluation Metrics: Binary Classification

## Receiver Operating Characteristic (ROC) Curve

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

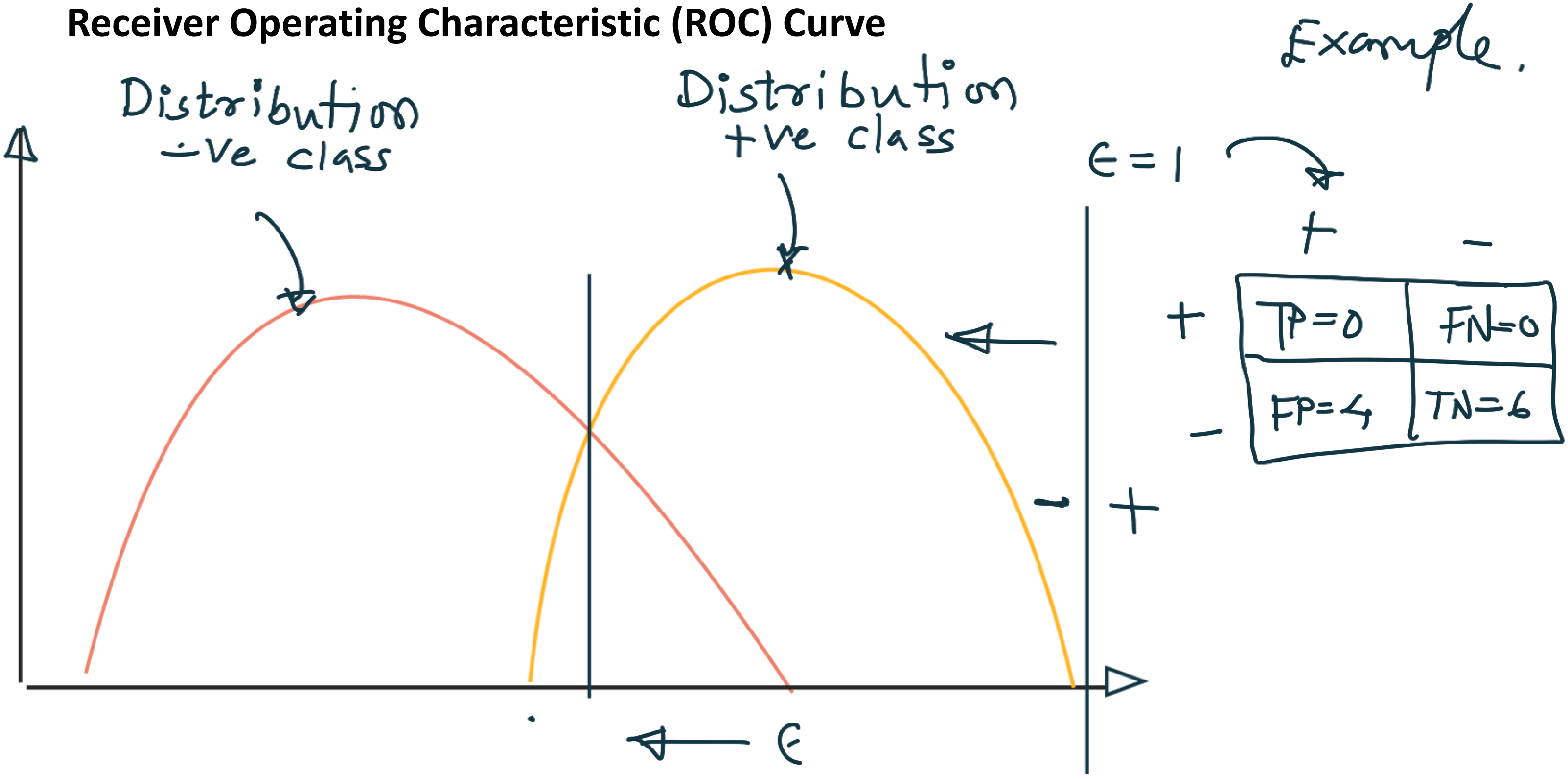
False Positive Rate = $\frac{FP}{FP+TN}$

| Points | Probability | Threshold | TPR | FPR |
|---|---|---|---|---|
| + | 0.9 | | 1/4 | 0 |
| + | 0.85 | | 2/4 | 0 |
| + | 0.6 | | 3/4 | 0 |
| - | 0.49 | | 3/4 | 1/6 |
| - | 0.4 | | 3/4 | 2/6 |
| - | 0.35 | | 3/4 | 3/6 |
| - | 0.34 | | 3/4 | 4/6 |
| + | 0.32 | | 1 | 4/6 |
| - | 0.2 | | 1 | 5/6 |
| - | 0.1 | | 1 | 1 |



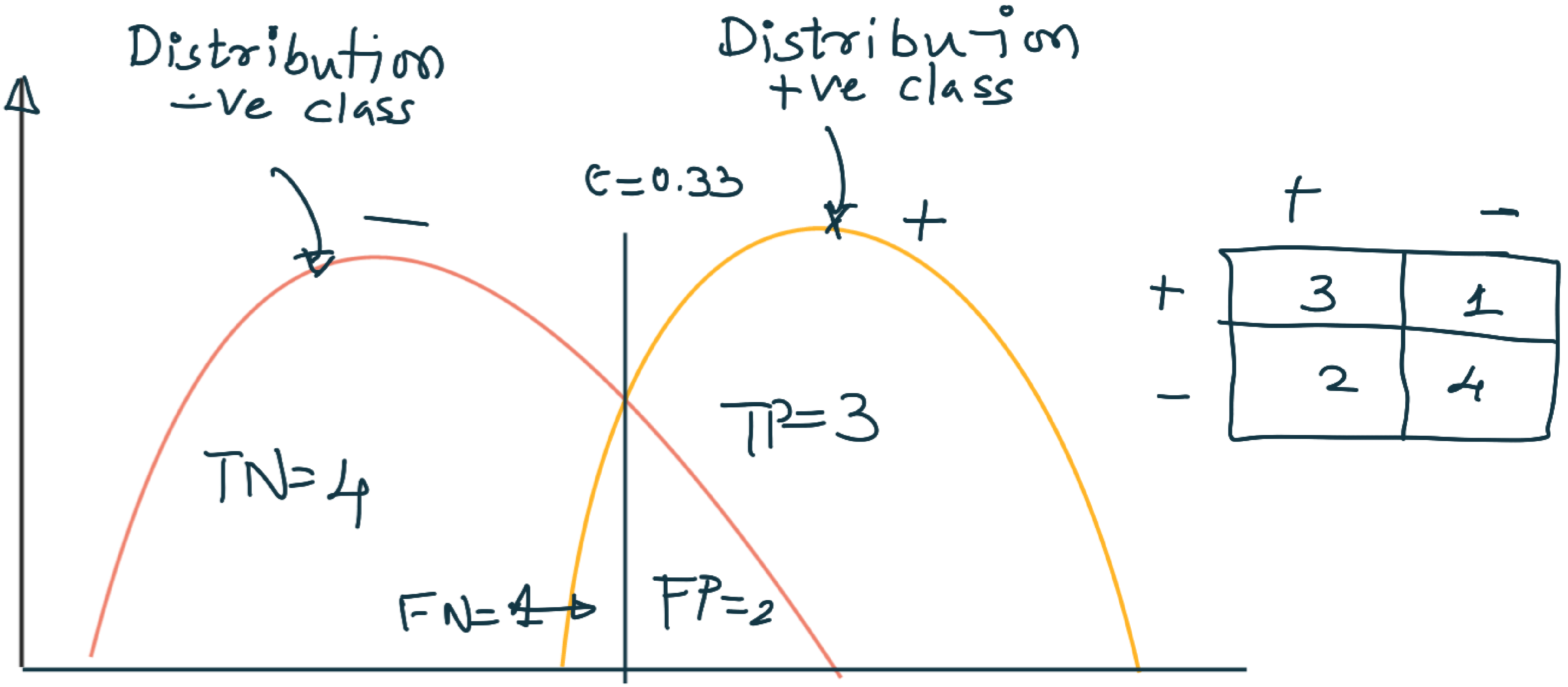
# Evaluation Metrics: Binary Classification

## Receiver Operating Characteristic (ROC) Curve



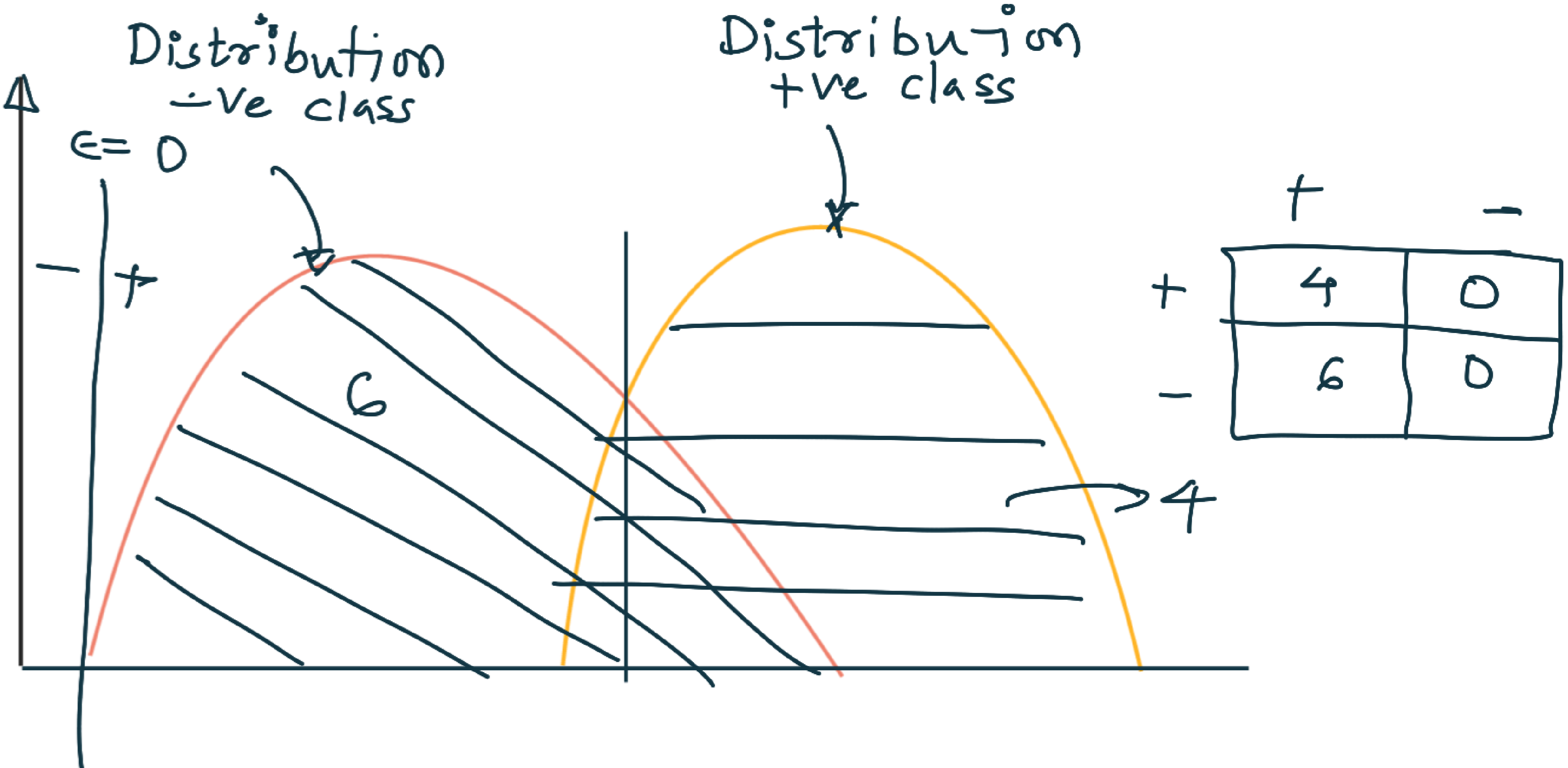
# Evaluation Metrics: Binary Classification

## Receiver Operating Characteristic (ROC) Curve



# Evaluation Matrix: Binary Classification

**Receiver Operating Characteristic (ROC) Curve**

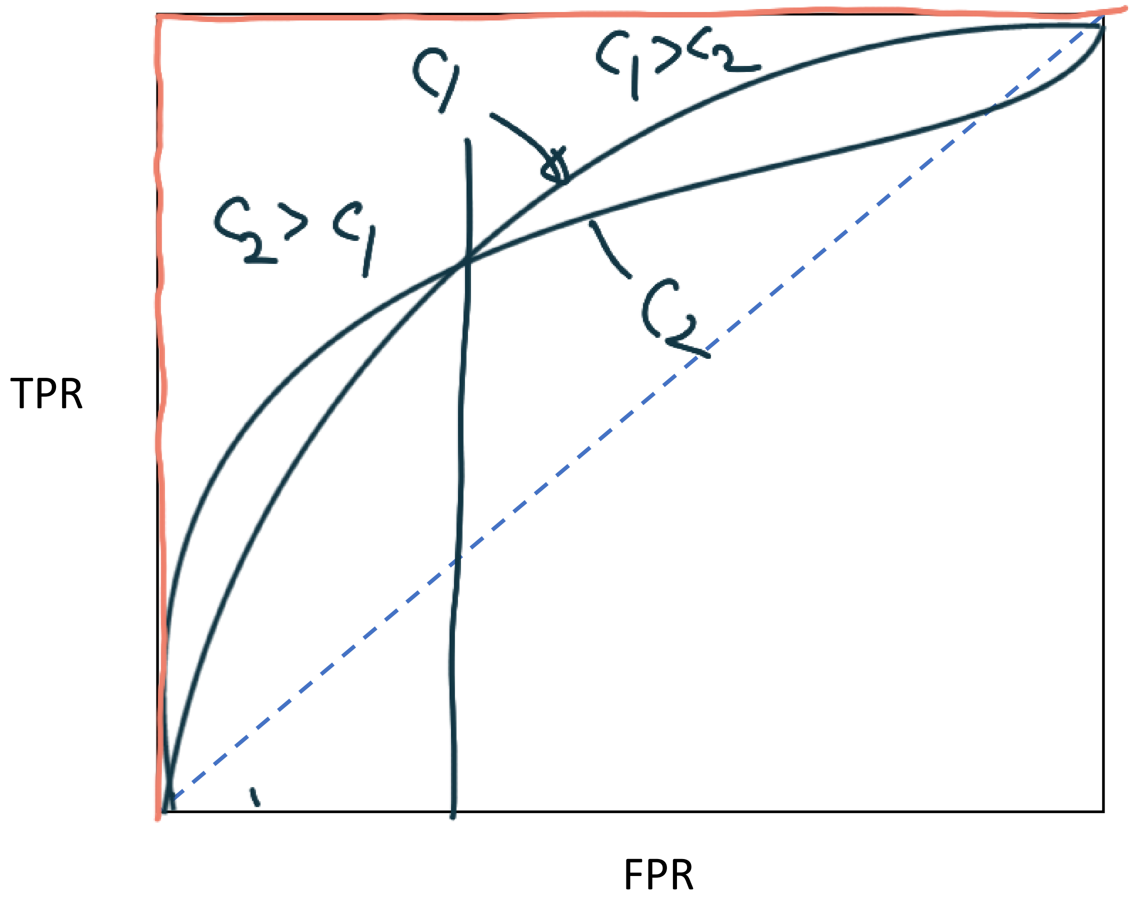


# Evaluation Metrics: Binary Classification

## Comparing Receiver Operating Characteristic (ROC) Curves

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

False Positive Rate = $\frac{FP}{FP+TN}$



# Evaluation Metrics: Binary Classification

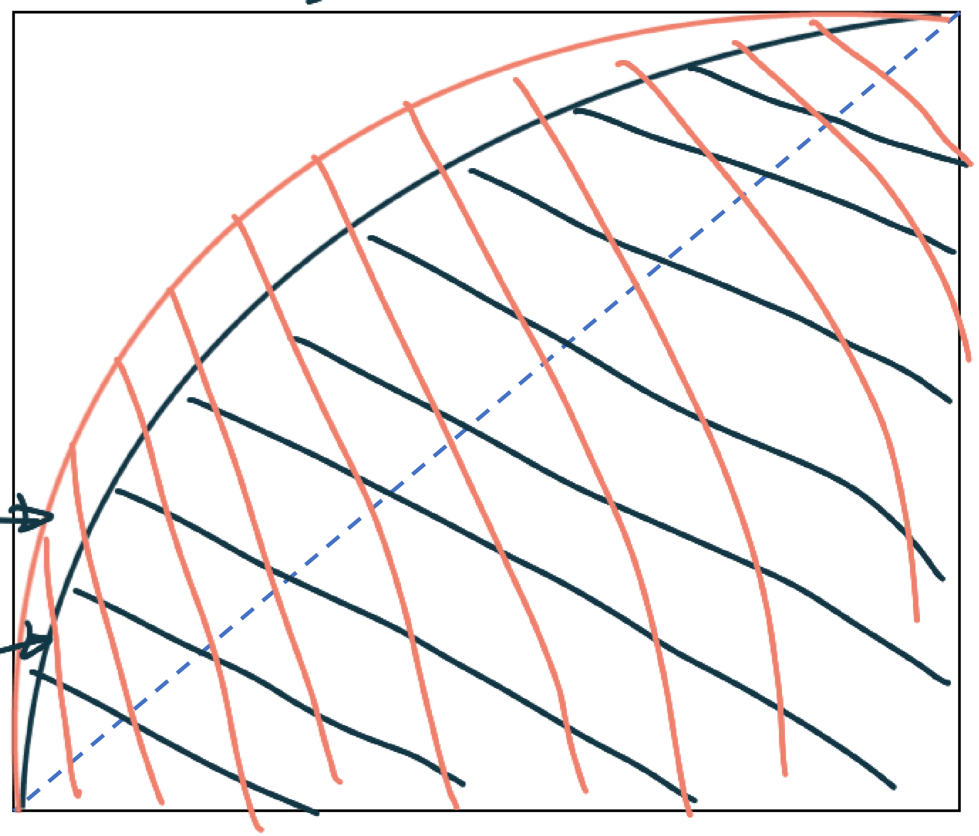
**Receiver Operating Characteristic (ROC) Curves**

Area under the curve (AUC)

$AUC(C_1) > AUC(C_2)$

$AUC(C_1)$

$AUC(C_2)$



# Evaluation Metrics: Binary Classification

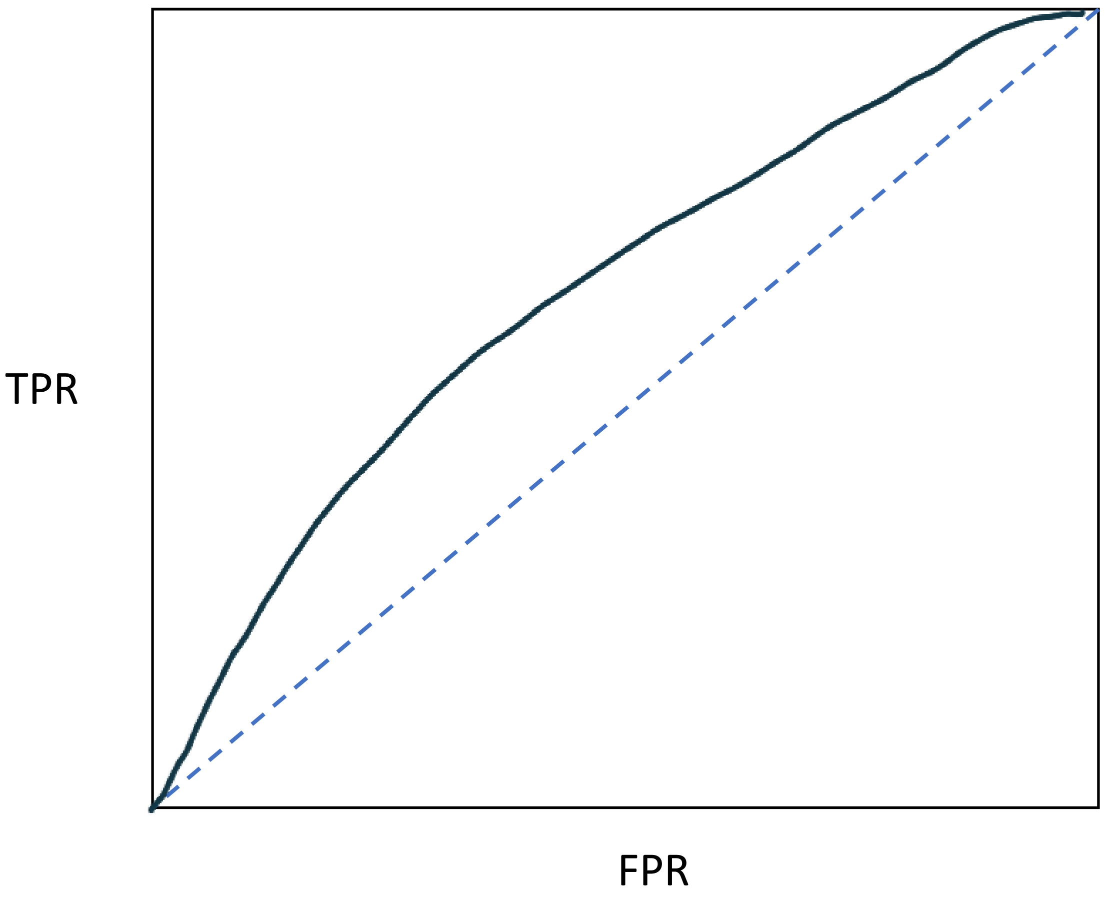
**Receiver Operating Characteristic (ROC) Curves** (class Imbalance)

Ratio −ve/+ve >>> 1

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

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

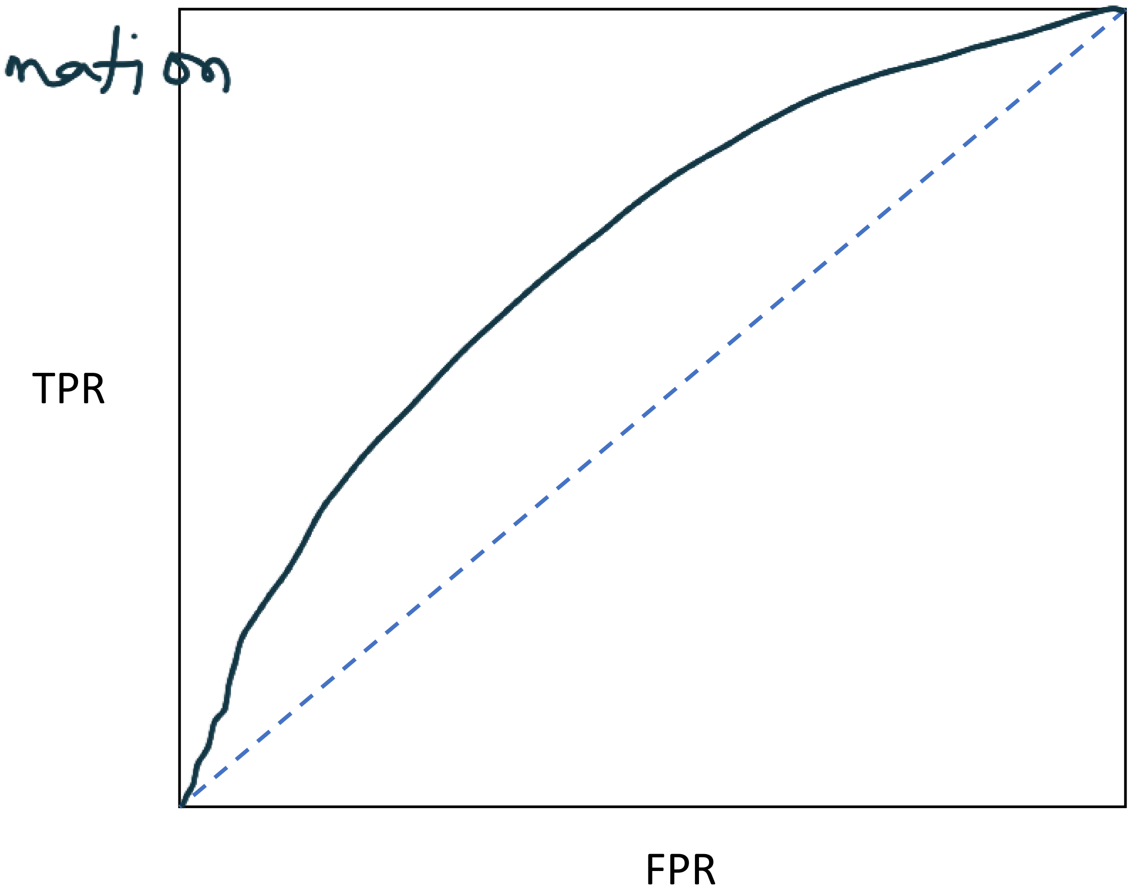
| | | P | | |
|---|---|---|---|---|
| | | + | − | |
| T | + | TP | FN | $\rightarrow$ TPR |
| | − | FP | TN | $\rightarrow$ FPR |



# Evaluation Metrics: Binary Classification

**Receiver Operating Characteristic (ROC) Curves**

- TPR & FPR use only a single row information
- Recall, F-measure use both row information
- ROC graphs insensitive to class imbalance.



# Evaluation Metrics: Binary Classification

## Receiver Operating Characteristic (ROC) Curves

(class Imbalance)

| | |
|---|---|
| 8 | 2 |
| 10 | 990 |

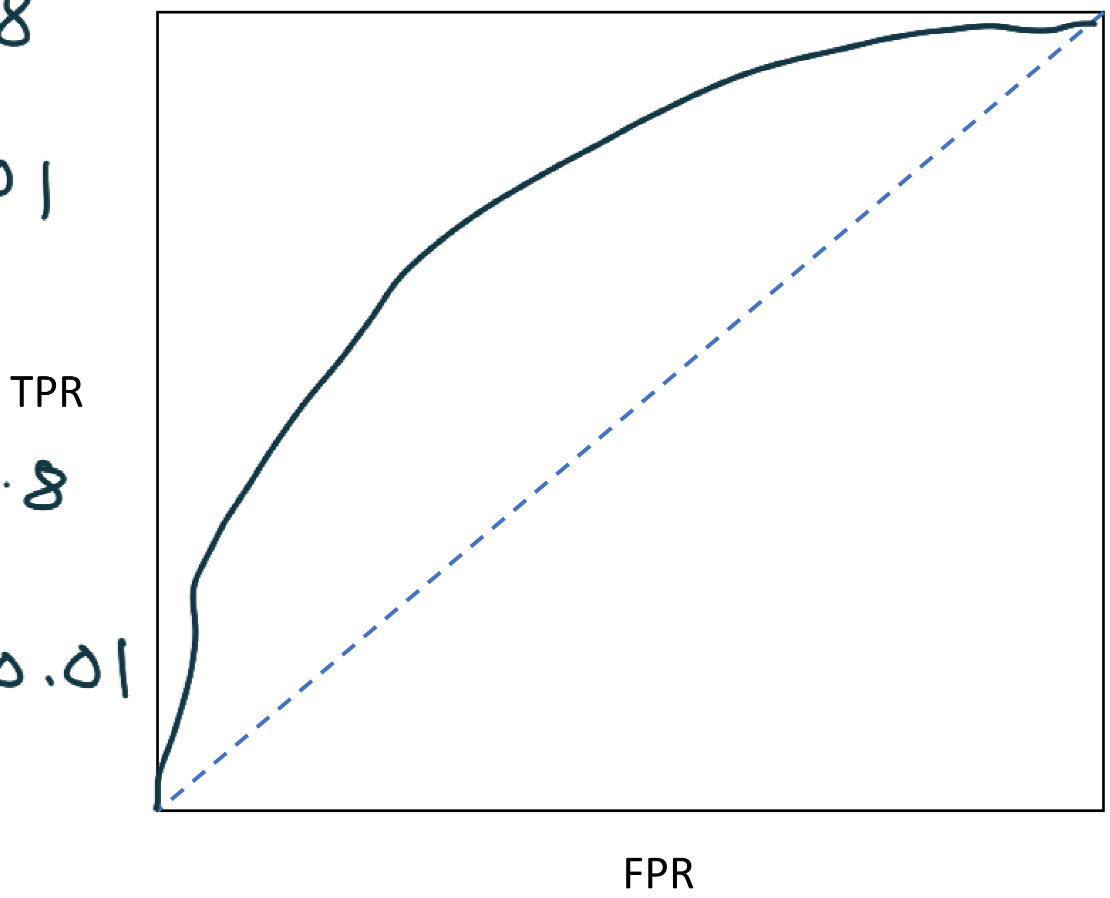
$$TPR = \frac{8}{10} = 0.8$$

$$FPR = \frac{10}{1000} = 0.01$$

| | |
|---|---|
| 16 | 4 |
| 2 | 198 |

$$TPR = \frac{16}{20} = 0.8$$

$$FPR = \frac{2}{200} = 0.01$$



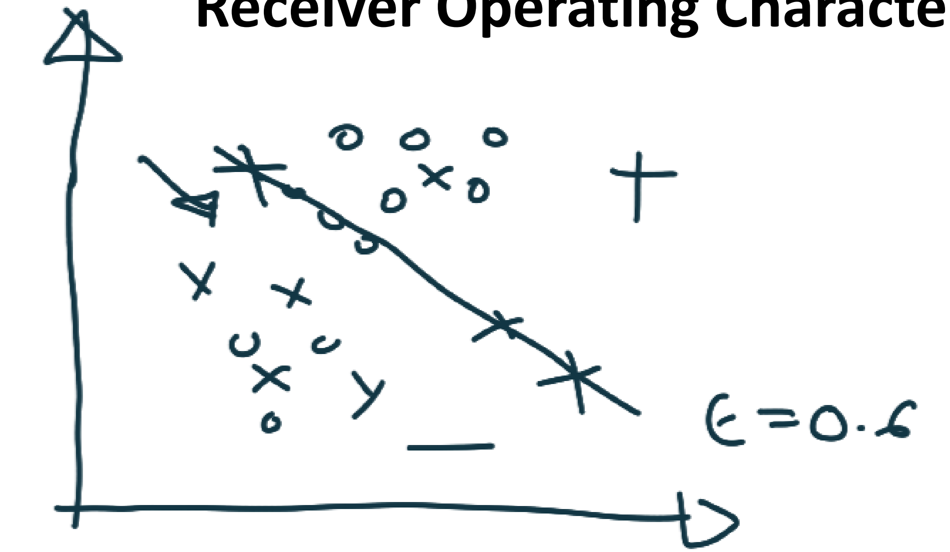
# Evaluation Metrics: Binary Classification

- class Imbalance affects on ROC curves – Negligible

- class Imbalance affects on Precision - Recall curves – significant.

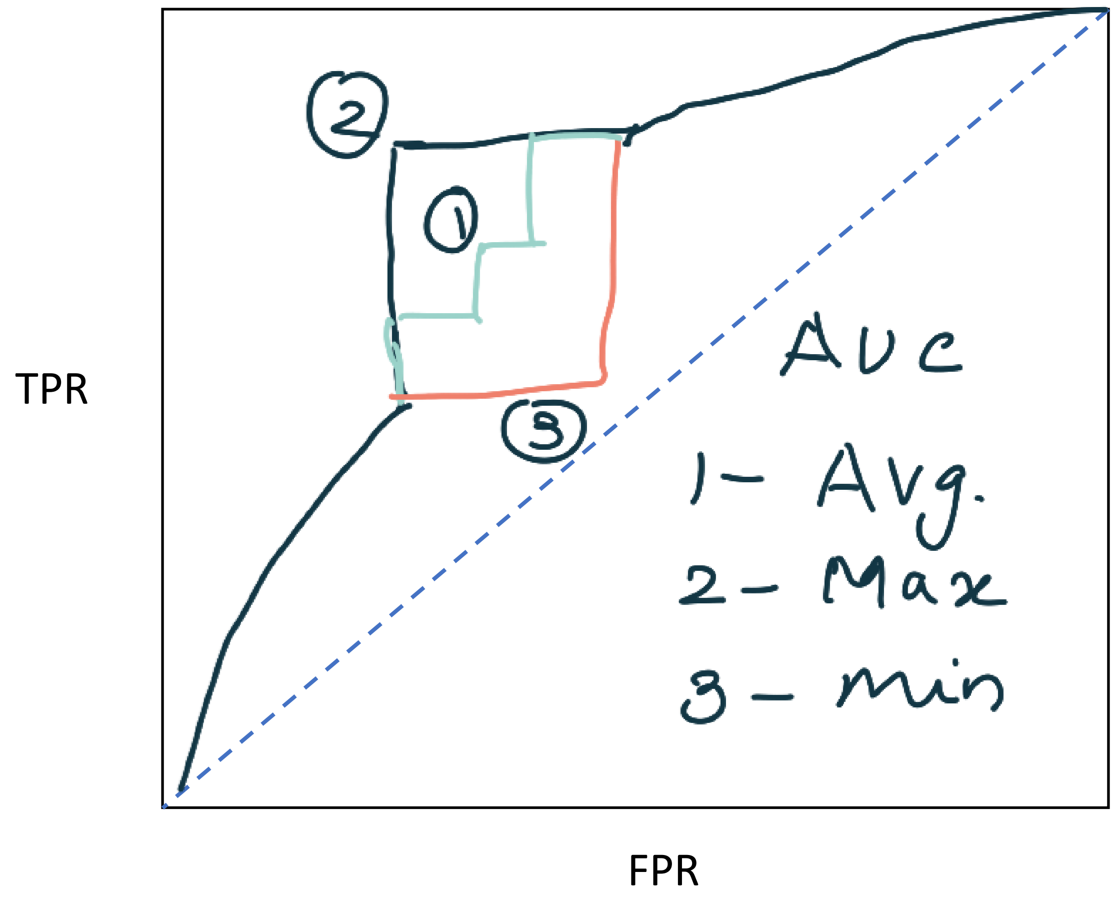
# Evaluation Metrics: Binary Classification

**Receiver Operating Characteristic (ROC) Curves** : Efficient Generation




Possible sortings

| (1) | (2) | (3) |
|---|---|---|
| + | + | – |
| – | + | – |
| + | + | – |
| – | – | + |
| + | – | + |
| – | – | + |



# Evaluation Metrics: Binary Classification

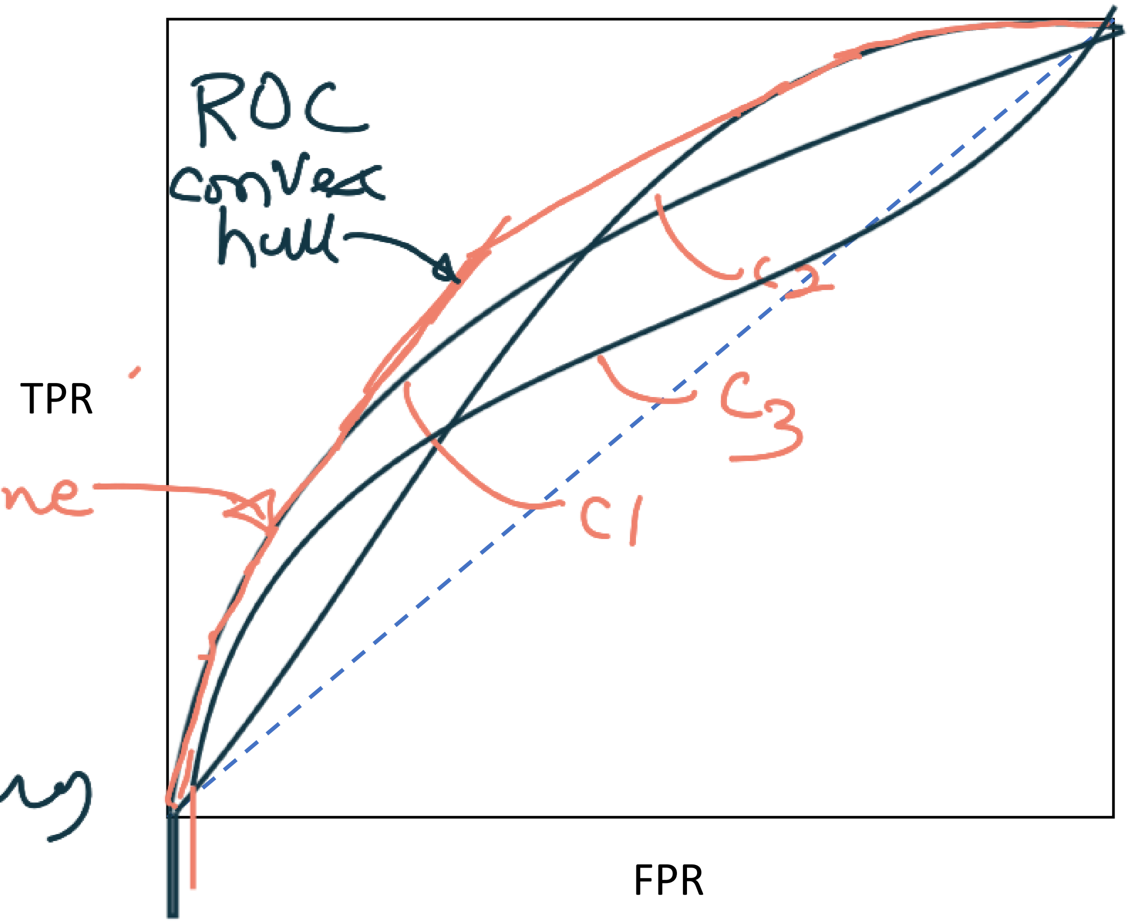
**Receiver Operating Characteristic (ROC) Curves** ROC convex hull

Two points in ROC curve

$$m = \frac{TPR_2 - TPR_1}{FPR_2 - FPR_1}$$

Iso-performance line

* Points on ROC convex hull give optimal classifiers ex. $C_1$ & $C_2$

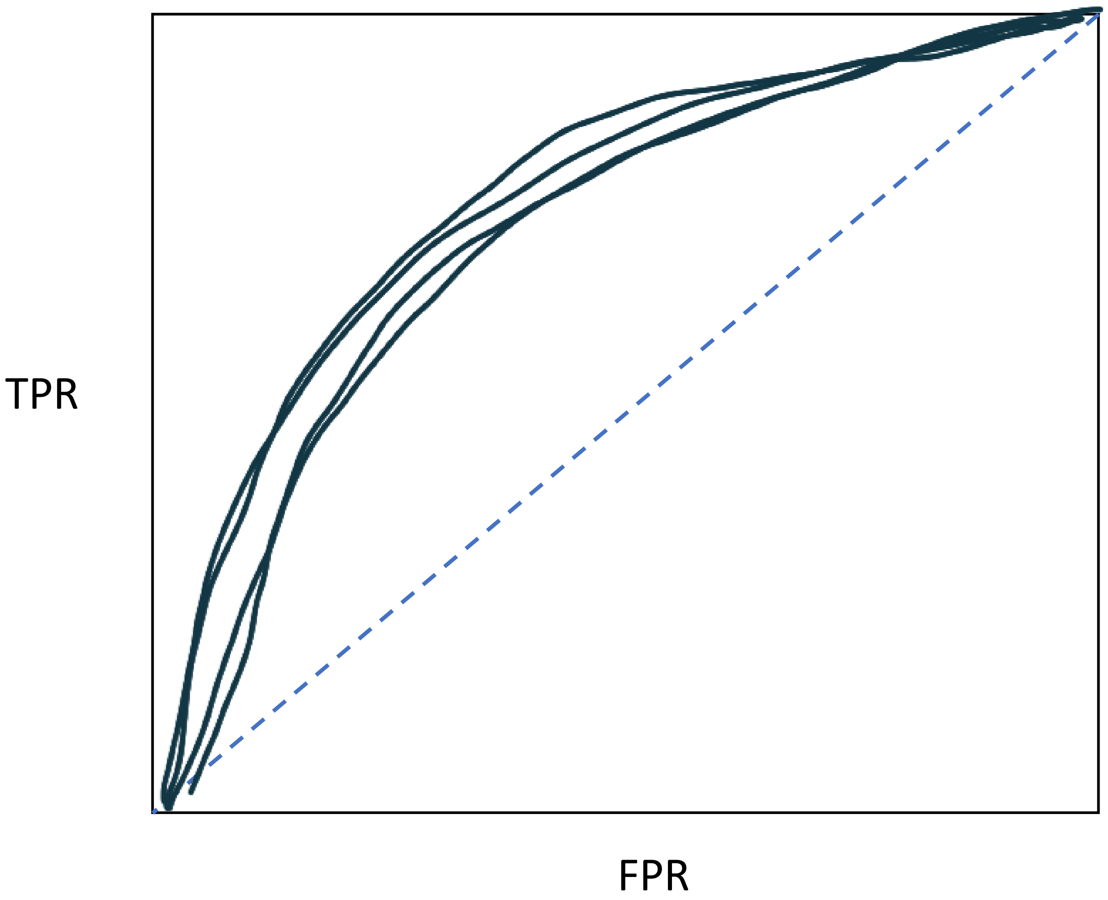


# Evaluation Metrics: Binary Classification

**Receiver Operating Characteristic (ROC) Curves** Test Data

* K-fold CV,
  Bootstrap,
  $T_1, T_2 \ldots T_K$

* Average ROC
  → classifier Performance



# Evaluation Metrics: Binary Classification

## Receiver Operating Characteristic (ROC) Curves

(vertical Averaged ROC)

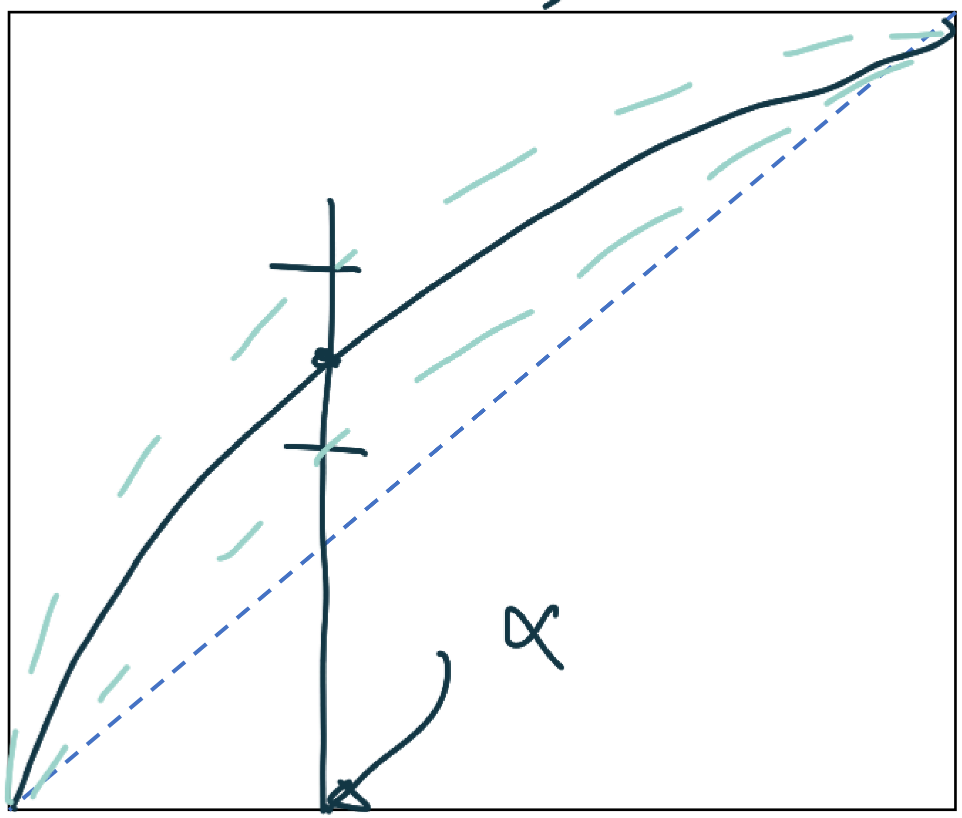
$$TPR = f_i(FPR)$$

$$= \text{mean}(FPR_\alpha)$$

$$TPR_{avg.}$$

$$= \frac{TPR^1_\alpha + TPR^2_\alpha + TPR^3_\alpha + TPR^4_\alpha}{4}$$

TPR




FPR

# Evaluation Metrics: Binary Classification

~~Precision Recall~~ ROC **Curve** (Threshold Averaging)

FPR → Not under control.

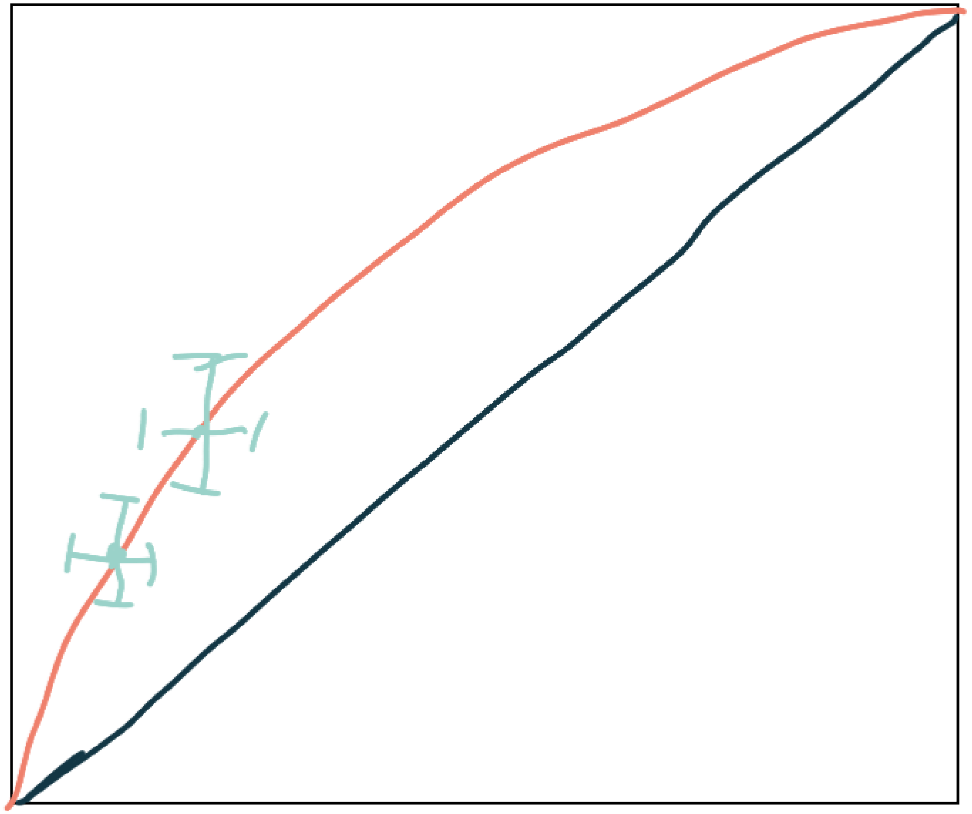
→ C :- Threshold Avg.

$T_1$ – $(TPR, FPR)_1$

$T_2$ – $(TPR, FPR)_2$

$T_3$ – $(TPR, FPR)_3$

$T_4$ – $(TPR, FPR)_4$



# Evaluation Metrics: ~~Multi~~-Class Classification

Binary

**Metrics**

* $TPR^{\epsilon}_{Avg.} = \text{mean}(TPR^{\epsilon}_1, \ldots, TPR^{\epsilon}_K)$
* $FPR^{\epsilon}_{Avg.} = \text{mean}(FPR^{\epsilon}_1, \ldots, FPR^{\epsilon}_K)$
* $TPR^{\epsilon}_{\sigma} = \text{std.}(TPR^{\epsilon}_1, \ldots, TPR^{\epsilon}_K)$
* $FPR^{\epsilon}_{\sigma} = \text{std.}(FPR^{\epsilon}_1, \ldots, FPR^{\epsilon}_K)$

# Evaluation Metrics: Multi-Class Classification

**Metrics** $\rightarrow$ Binary class to multi-class

Three class Problem: 0, 1, 2

Predicted

| True \ Predicted | 0 | 1 | 2 |
|---|---|---|---|
| 0 | $TP_0$ | $E_{01}$ | $E_{02}$ |
| 1 | $E_{10}$ | $TP_1$ | $E_{12}$ |
| 2 | $E_{20}$ | $E_{21}$ | $TP_2$ |

E: Misclassification

$0^{th}$ class predicted as $2^{nd}$ class

one / Rest approach

$2 \rightarrow +$, $\quad 0, 1 \rightarrow -$

| | |
|---|---|
| TP | FN |
| FP | TN |

$\rightarrow$

| | |
|---|---|
| $TP_2$ | $E_{20} + E_{21}$ |
| $E_{02} + E_{12}$ | $TP_0 + TP_1$ |

# Evaluation Metrics: Multi-Class Classification

**Metrics**

| | |
|---|---|
| $TP_2$ | $E_{20}+E_{21}$ |
| $E_{02}+E_{12}$ | $TP_0+TP_1$ |

$\rightarrow (Precision)_2 = \dfrac{TP_2}{TP_2 + E_{02} + E_{12}}$

$\rightarrow (Recall)_2 = \dfrac{TP_2}{TP_2 + E_{20} + E_{21}}$

$1 \rightarrow +, \; 0,2 \rightarrow -$

$(Precision)_1 = TP_1/(TP_1 + E_{01} + E_{21})$

$(Recall)_1 = TP_1/(TP_1 + E_{10} + E_{12})$

$0 \rightarrow +, \; 1,2 \rightarrow -$

$(Precision)_0 = TP_0/(TP_0 + E_{10} + E_{20})$

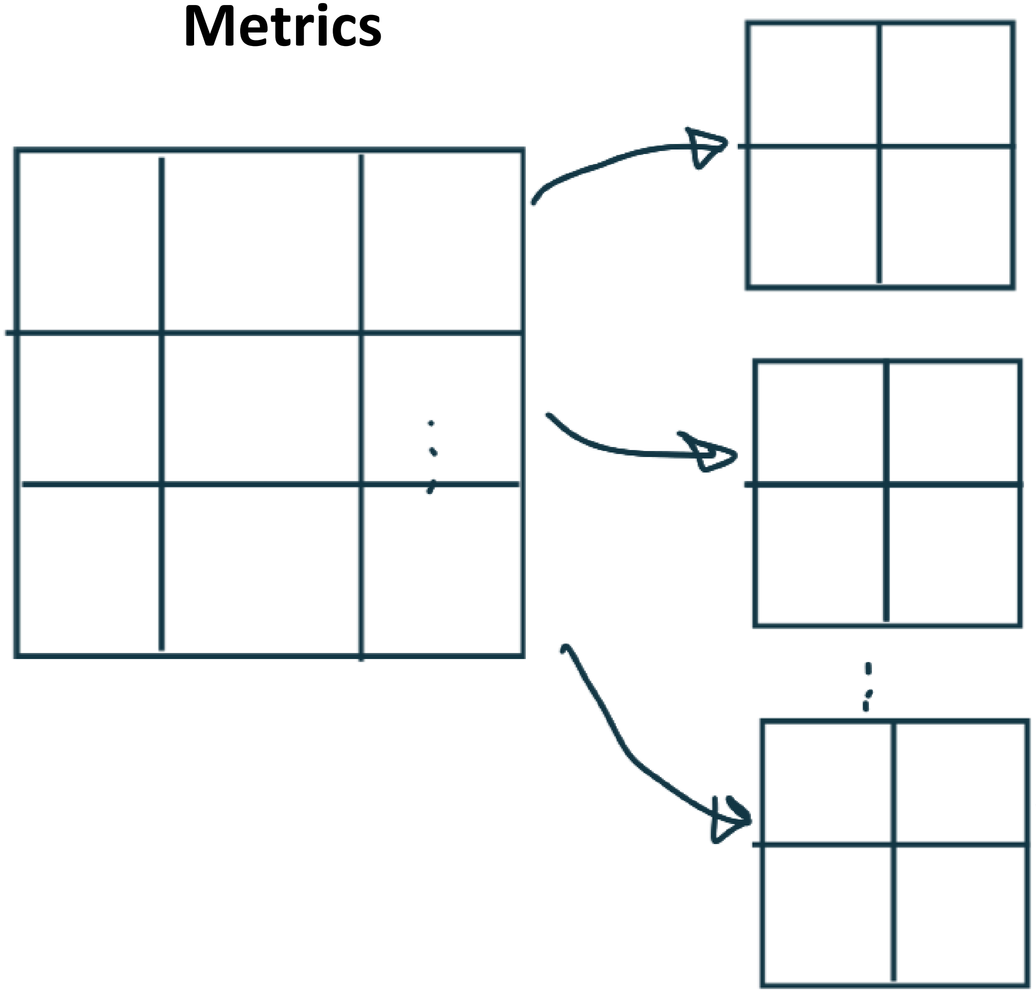
$(Recall)_0 = TP_0/(TP_0 + E_{01} + E_{02})$

Overall Precision/Recall $\rightarrow P_{ov} = \sum_{i=0}^{2} (Precision)_i/3, \; R_{ov} = \sum_{i=0}^{2} (Recall)_i/3$

# Evaluation Metrics: Multi-Class Classification

Macro Averaging

**Metrics**



$\rightarrow PC_1, RC_1$

$- PC_2, RC_2$

$\vdots$

$- PC_K, RC_K$

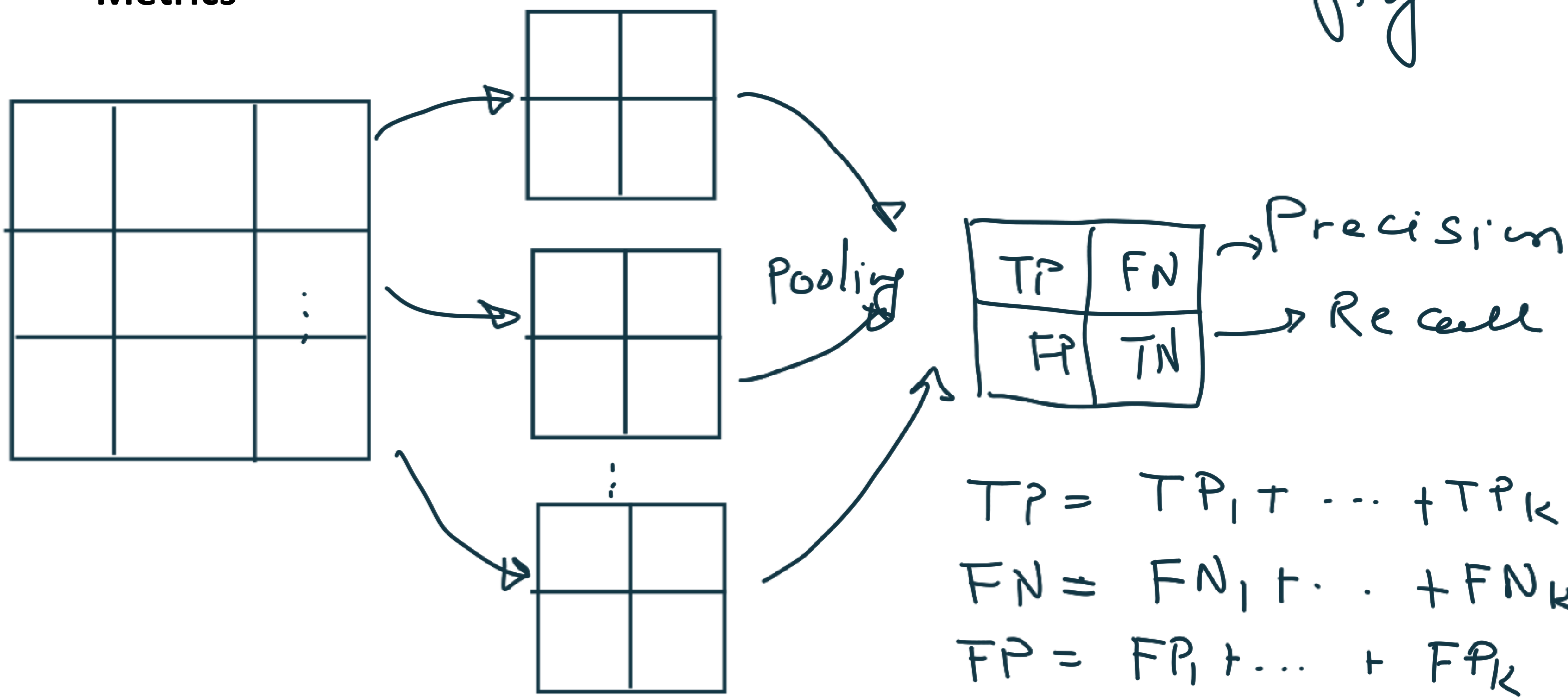
$$P_{overall} = \frac{\sum_{i=1}^{K} PC_i}{K}$$

$$R_{overall} = \frac{\sum_{i=1}^{K} RC_i}{K}$$

# Evaluation Metrics: Multi-Class Classification

**Metrics**

Micro Averaging




$$TP = TP_1 + \cdots + TP_K$$

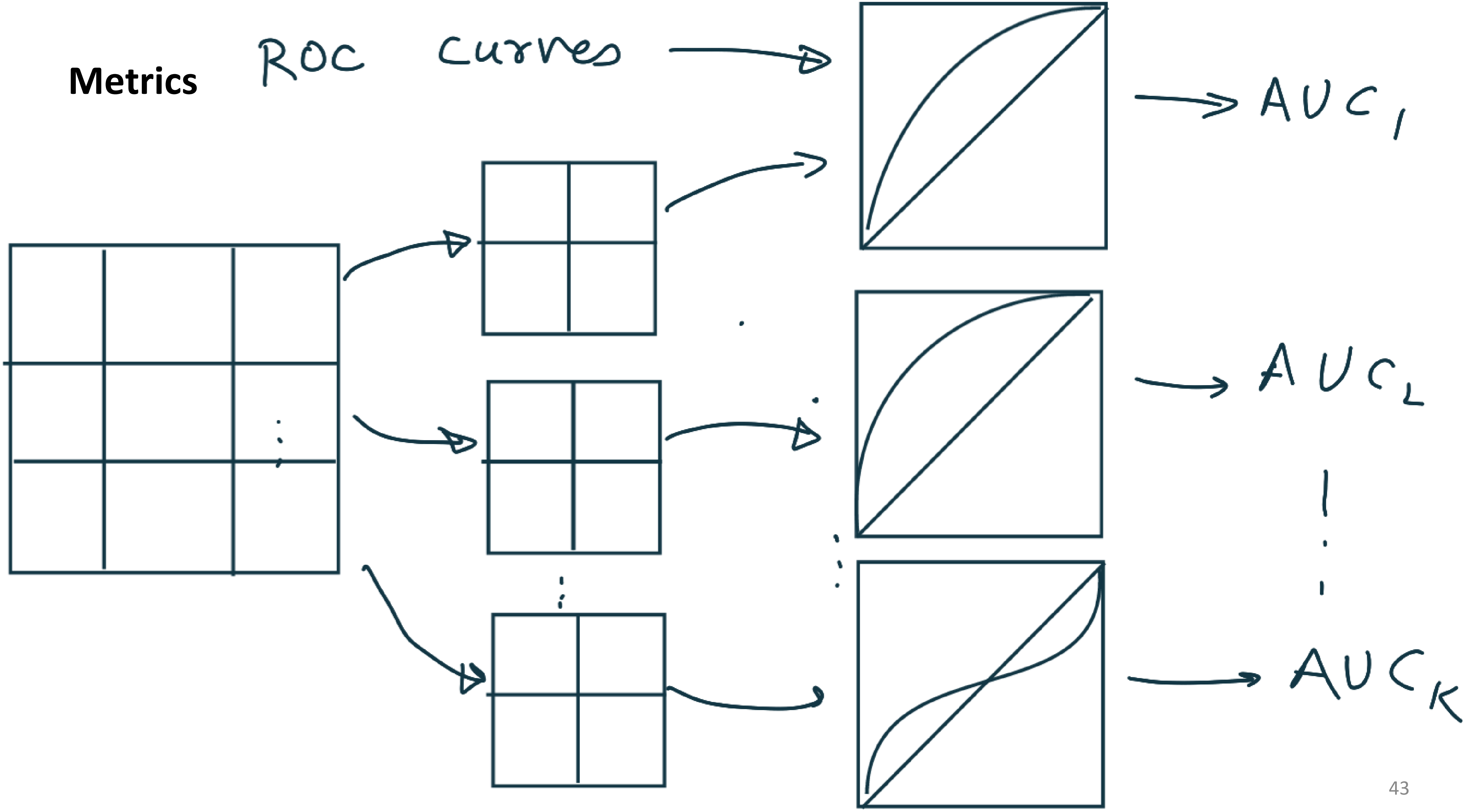
$$FN = FN_1 + \cdots + FN_K$$

$$FP = FP_1 + \cdots + FP_K$$

$$TN = TN_1 + \cdots + TN_K$$

# Evaluation Metrics: Multi-Class Classification

**Metrics**



# Evaluation Metrics: Multi-Class Classification

**Metrics**

* $$AUC_{overall} = \frac{AUC_1 + \ldots \quad + AUC_K}{K}$$

$$= \sum_{i=1}^{K} \frac{AUC_i}{K}$$

# Evaluation Metrics: Multi-Class Classification

**Metrics**

* $AUC_{overall}$

$$= \sum_{i=1}^{K} AUC_i \, p(i)$$

$p(i) \rightarrow$ prevalenle (or fraction) of the ith class in the training data.

# Evaluation Metrics: Multi-Class Classification

**Metrics**

Third Approach

* combination of binary class problems.
* K - multiclass problems
  ↳ $K(K-1)/2$ binary class cominations.
* $$AUC_{overall} = \frac{2}{K(K-1)} \sum_{\substack{\{i,j\} \in C \\ i \neq j}} AUC(i,j)$$

↳ confusion matrix for multiclass problem.

# Evaluation Metrics: Multi-Class Classification

**Metrics**

summary

Accuracy

Precision

Recall

F1-measure

AUC / AUC overall

Exploratory

ROC curves

Precision - Recall curves

# References:

1. Tom Fawcett, An Introduction to ROC Analysis, Pattern Recognition Letters, 2006, 861-874
2. Alaa Tharwat, Classification Assessment Methods, Applied Computing and Informatics Vol. 17 No. 1, 2021 pp. 168-192