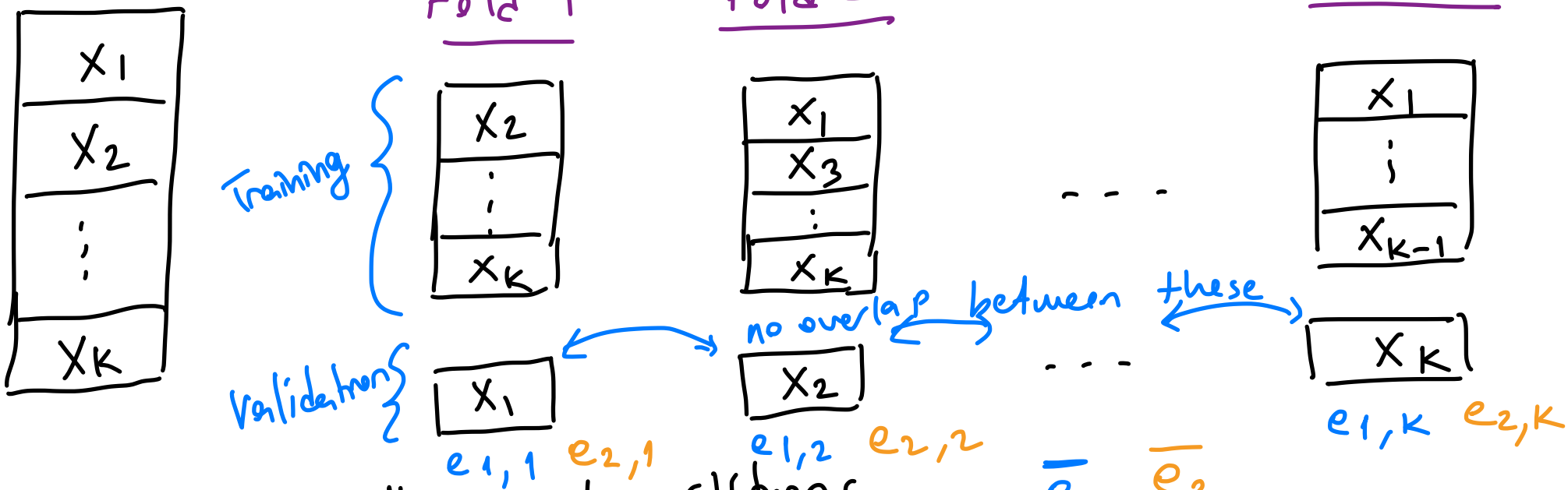
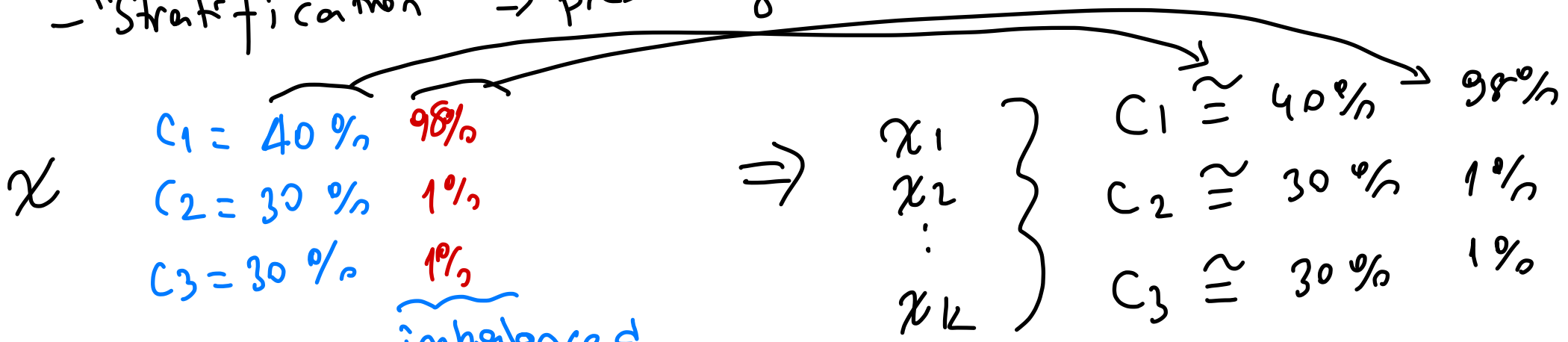
CROSS-VALIDATION METHODS

A1 A2

(A) K-Fold Cross-Validation



- K almost equally sized partitions
- "Stratification" ⇒ preserving class ratios in different partitions



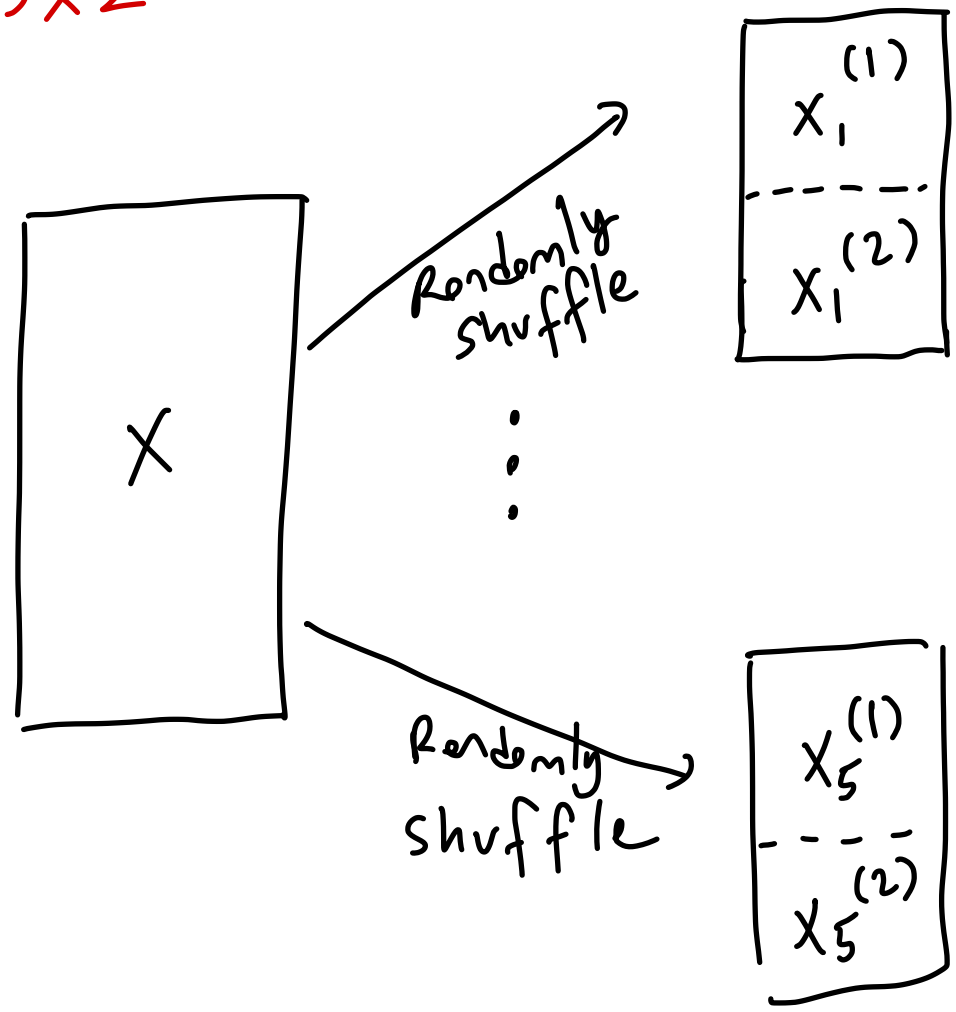
(B) LEAVE-ONE-OUT (LOO) CROSS-VALIDATION

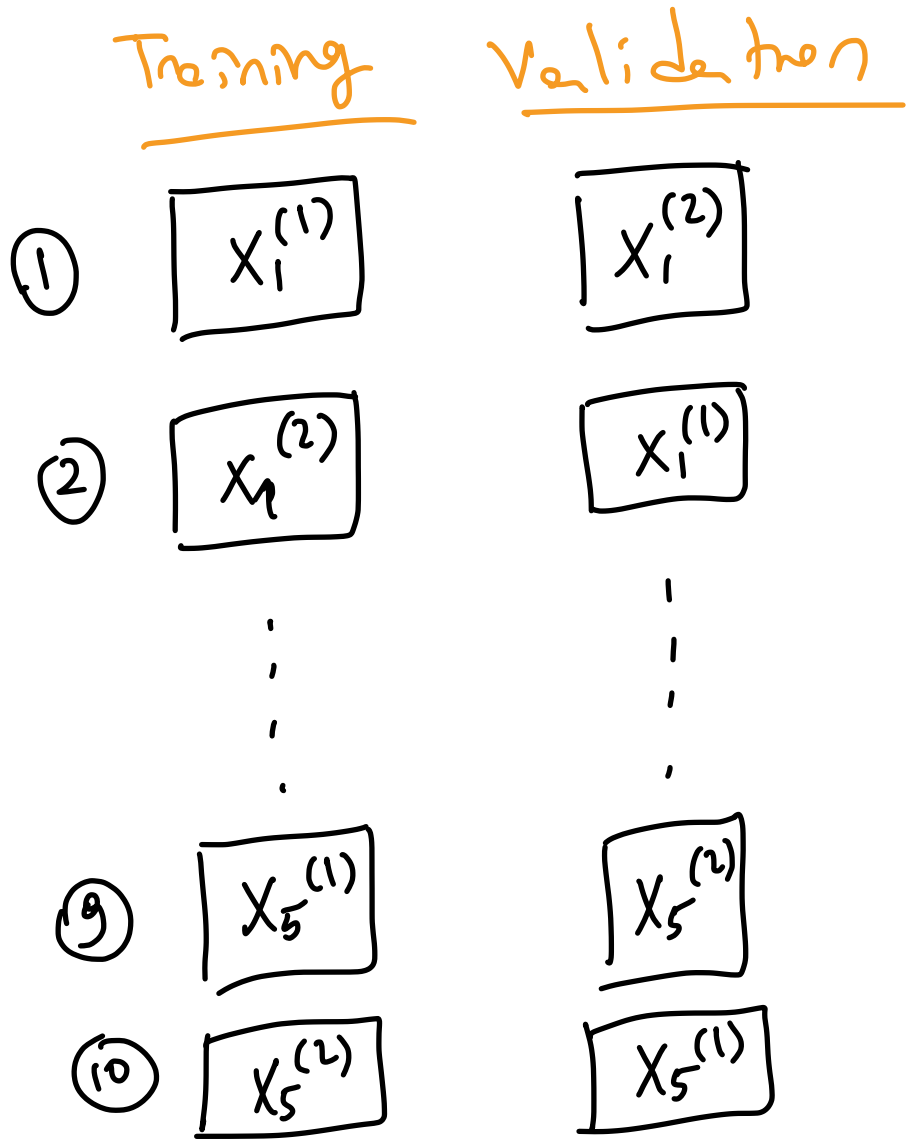
⇒ K-fold cross-validation when $K = N$

⇒ If N is very small, this maximizes the training set size

⇒ Biomedical applications

(C) 5x2 CROSS-VALIDATION





MEASURING CLASSIFIER PERFORMANCE

$$\text{Accuracy} = \frac{\#\text{ of correct predictions}}{\#\text{ of predictions}}$$

$$\text{Misclassification Error} = \frac{\#\text{ of incorrect predictions}}{\#\text{ of predictions}}$$

$$\frac{np.sum(np.diag(C))}{np.sum(C)} \quad (\text{diag} \rightarrow tp + tn)$$

$$1 - [\ ]$$

$$\chi_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}} \qquad x_i \in \mathbb{R}^D$$

$$\chi_{test} = \{x_i\}_{i=1}^{N_{test}} \qquad y_i \Rightarrow \text{unknown}$$

$y_i \in \{-1, +1\}$

$y_i \in \{0, 1\}$

$y_i \in \{+, -\}$

Truth \ Predicted	+	−
+	tp	fn
−	fp	tn

$p \rightarrow$ # of positives

$n \rightarrow$ # of negatives

$N \rightarrow$ # of data points

$p' \rightarrow$ # of positive predictions

$n' \rightarrow$ # of negative predictions

tp = true positive

tn = true negative

fp = false positive

fn = false negative

$$\text{Accuracy} = \frac{tp + tn}{tp + fn + fp + tn} = \frac{tp + tn}{N}$$

$$\text{Misclassification Error} = \frac{fp + fn}{N}$$

$tp\ rate = tp/p$ ⟷ $recall = tp/p$

$fp\ rate = fp/n$ ⟷ $sensitivity = tp/p$

$precision = tp/p'$ ⟷ $specificity = tn/n$

Imbalanced dataset $\begin{bmatrix} 98\% & + \\ 2\% & - \end{bmatrix}$

predicts (+) all the time

I \ P	+	−
+	98	0
−	2	0

Accuracy = 98/100

Precision = 98/100

Recall/Sens = 98/98 = 1

Specificity = 0/2 = 0

I \ P	+	−
+	95	3
−	0	2

Accuracy = 97/100

Precision = 95/95

Recall/Sens. = 95/98

Specificity = 2/2

RECEIVER OPERATING CHARACTERISTICS (ROC) CURVE

$P(+\mid x)$	0.4	0.6	0.7	0.5	0.9	0.3	0.2	0.1	0.8
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9

Step #1 Sort these probabilities/scores in increasing order.

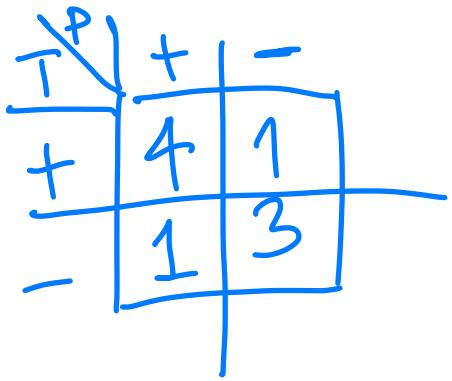
0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
x_8	x_7	x_6	x_1	x_4	x_2	x_3	x_9	x_5
−	+	−	−	+	+	−	+	+

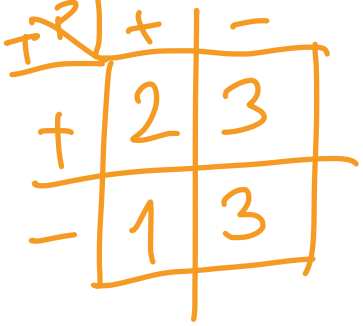
Step #2 Look at their correct labels.

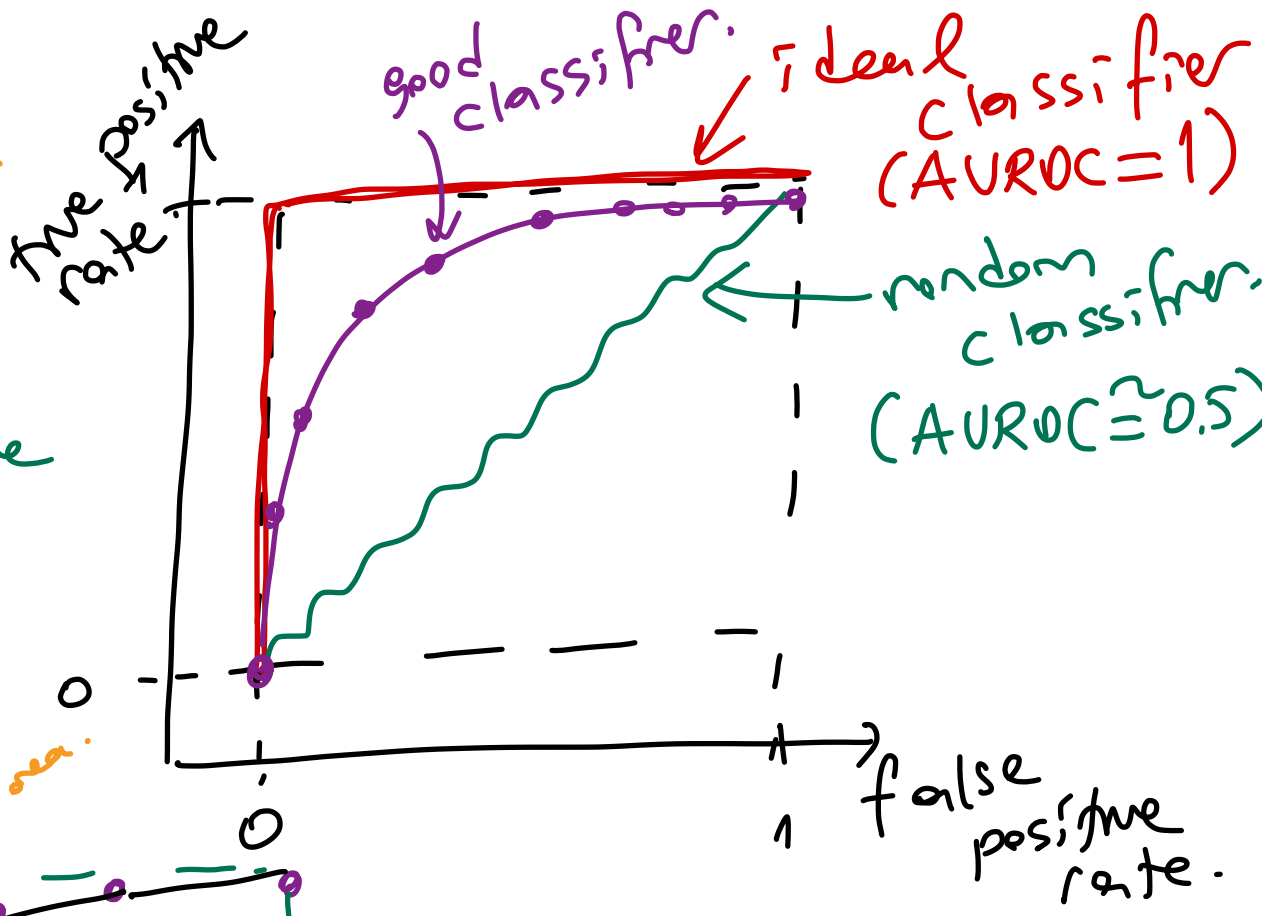
Ideal → − − − − + + + + +

Step #3 Calculate performance statistics with different thresholds.

Threshold 0.45	$-_{tn}$ ✓	$-_{fn}$ ✗	$-_{tn}$ ✓	$-_{tn}$ ✓	$+_{tp}$ ✓	$+_{tp}$ ✓	$+_{fp}$ ✗	$+_{tp}$ ✓	$+_{tp}$ ✓
Threshold 0.65	$-_{tn}$ ✓	$-_{fn}$ ✗	$-_{tn}$ ✓	$-_{tn}$ ✓	$-_{fn}$ ✗	$-_{fn}$ ✗	$+_{fp}$ ✗	$+_{tp}$ ✓	$+_{tp}$ ✓

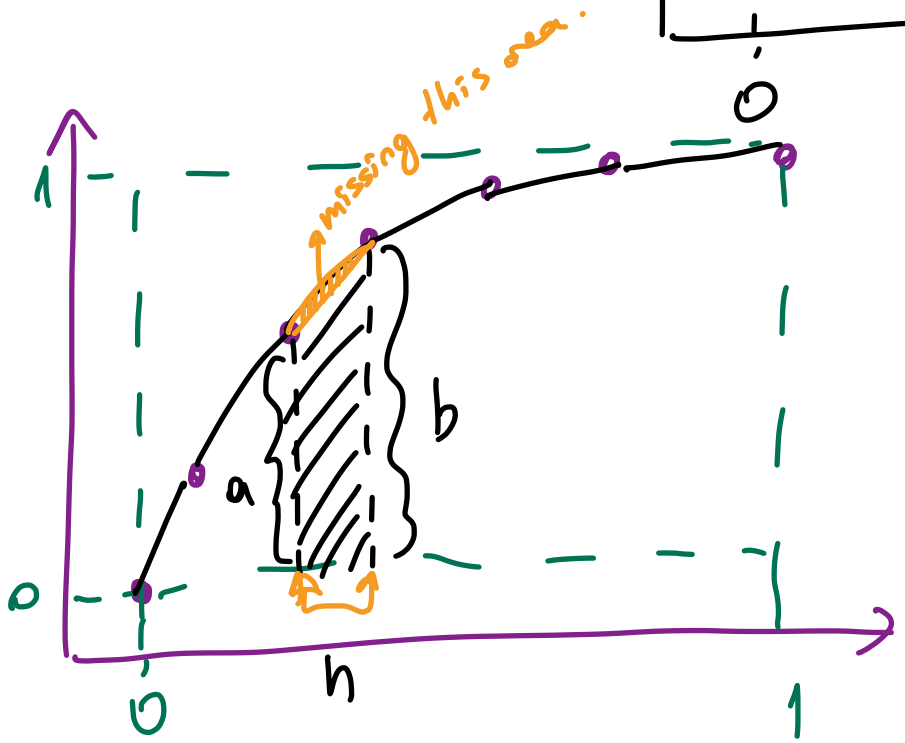






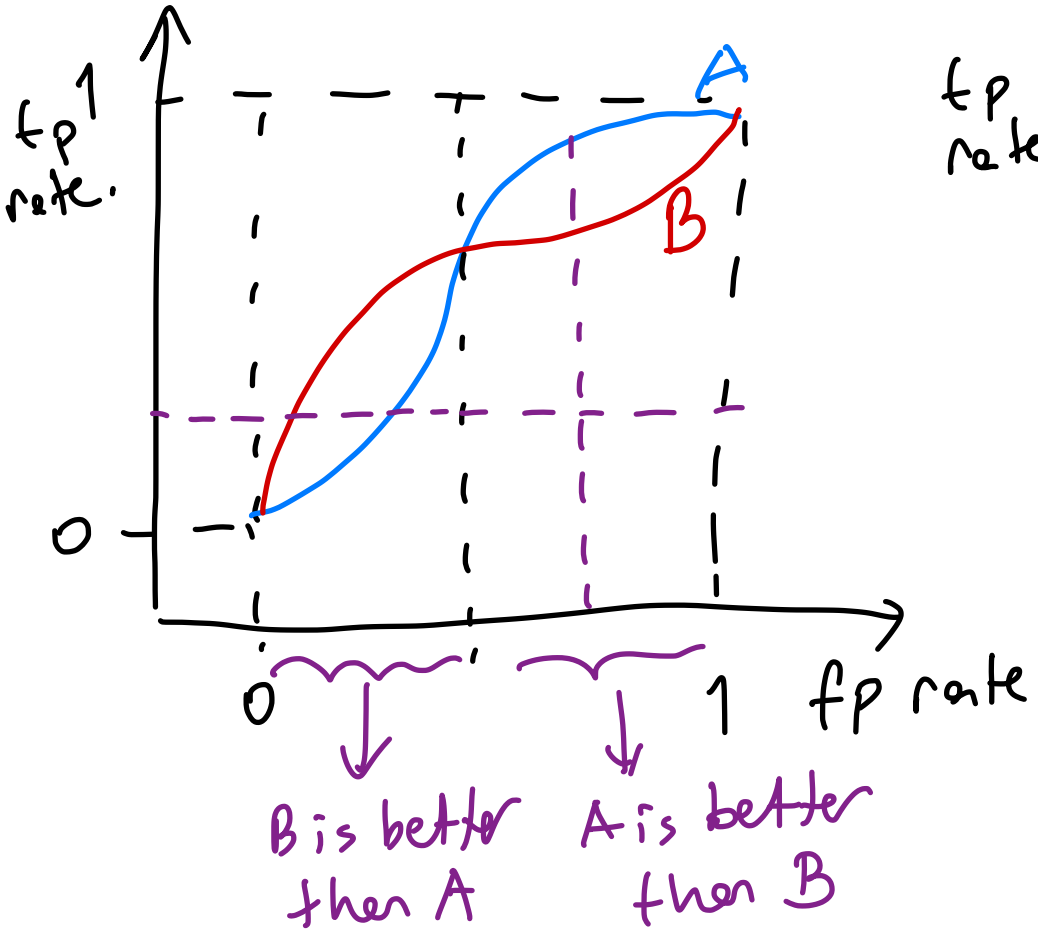
Area under the ROC curve

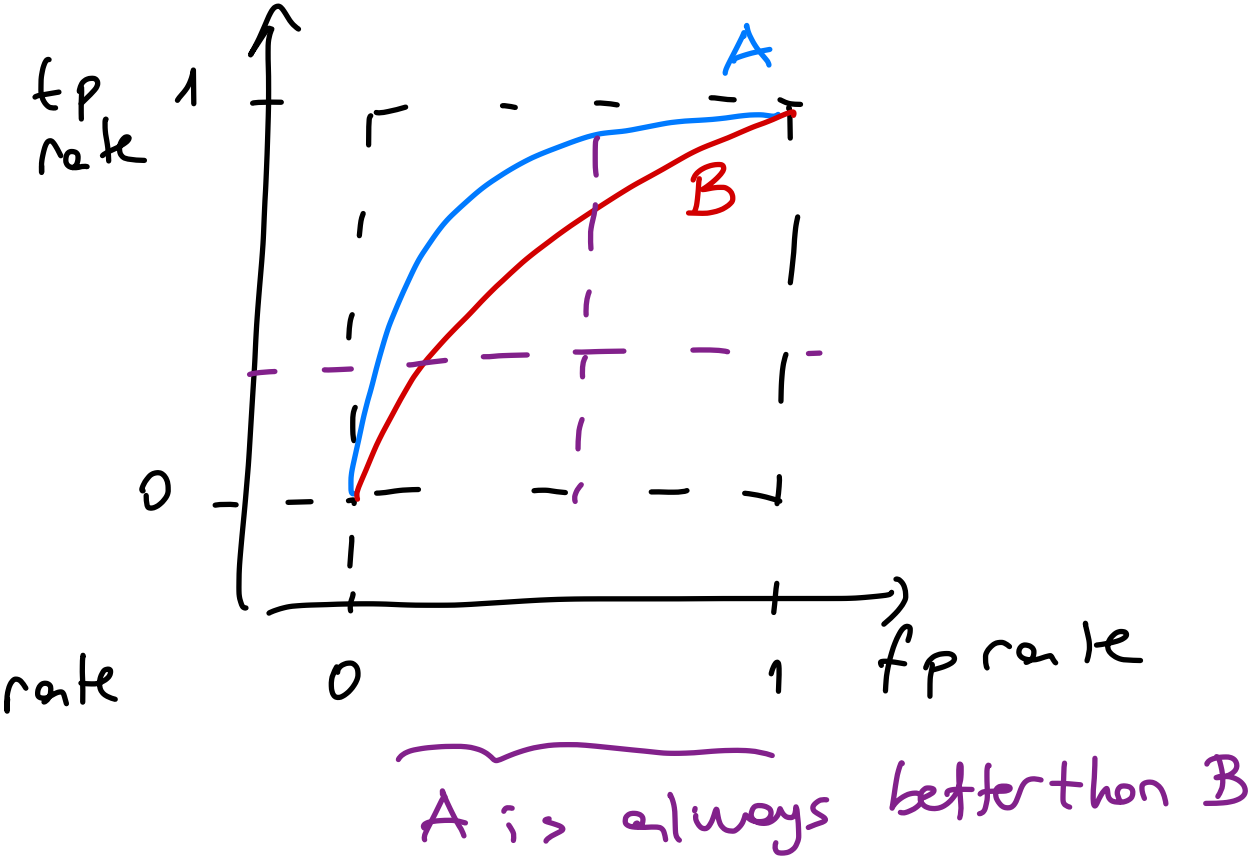
$$\int_0^1 f(x)\,dx$$



"trapezoid rule"

$$\frac{(a+b)\cdot h}{2}$$


tp rate
1
A
B
0
0
1
fp rate
B is better than A
A is better than B


tp rate
1
A
B
0
0
1
fp rate
A is always better than B