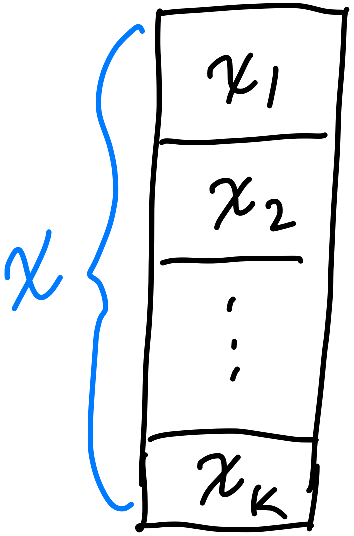
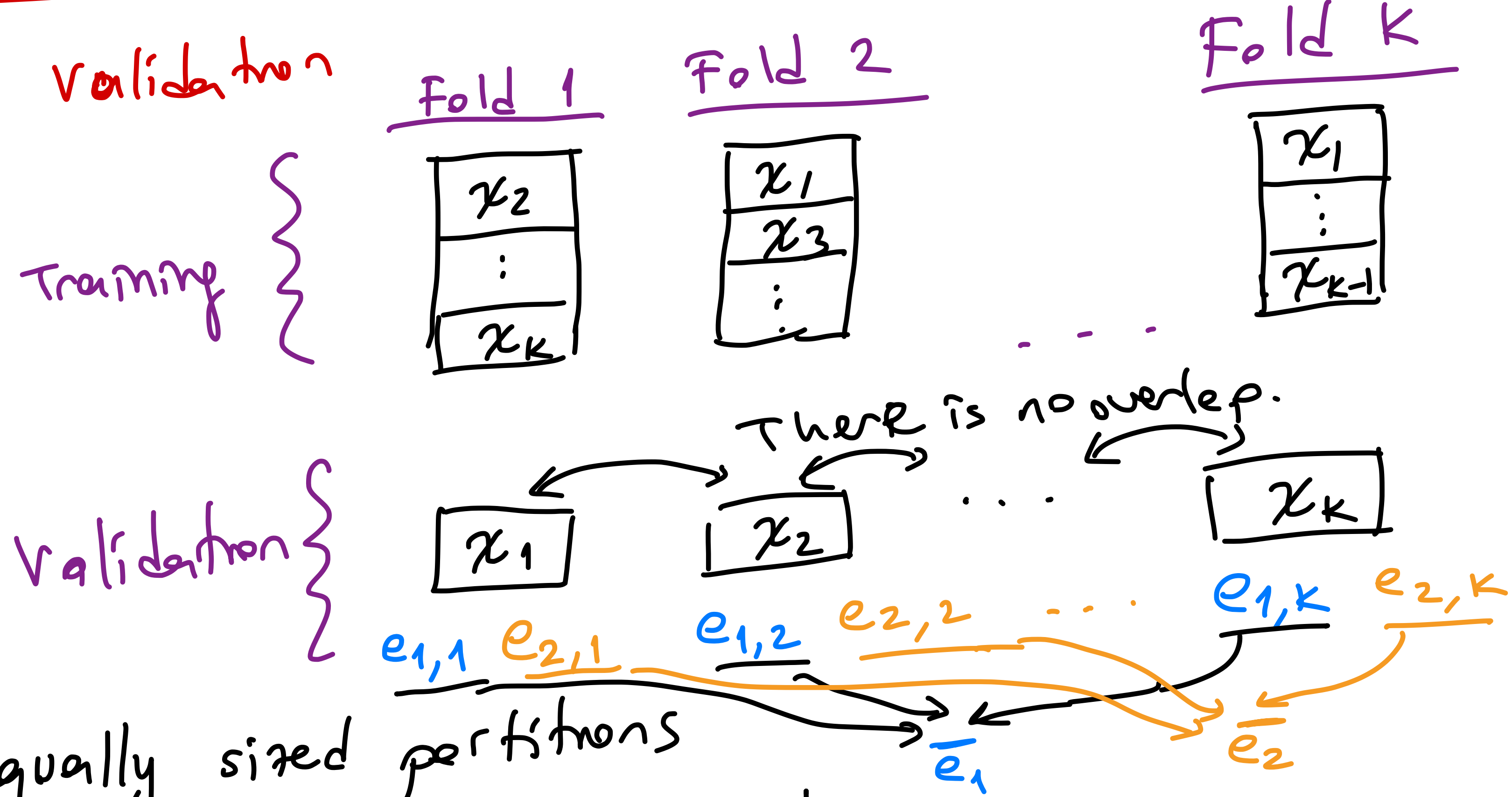
CROSS-VALIDATION METHODS

(A) K-Fold Cross validation





- K almost equally sized partitions
- "stratification" $\Rightarrow$ preserving class ratios in different partitions

$$\mathcal{X} \Rightarrow \begin{array}{l} C_1 = 40\% \\ C_2 = 30\% \\ C_3 = 30\% \end{array} \left\} \begin{array}{ll} 98\% & X_1 \\ 1\% & \vdots \\ 1\% & X_K \end{array} \right\} \begin{array}{ll} C_1 \cong 40\% & 98\% \\ C_2 \cong 30\% & 1\% \\ C_3 \cong 30\% & 1\% \end{array}$$

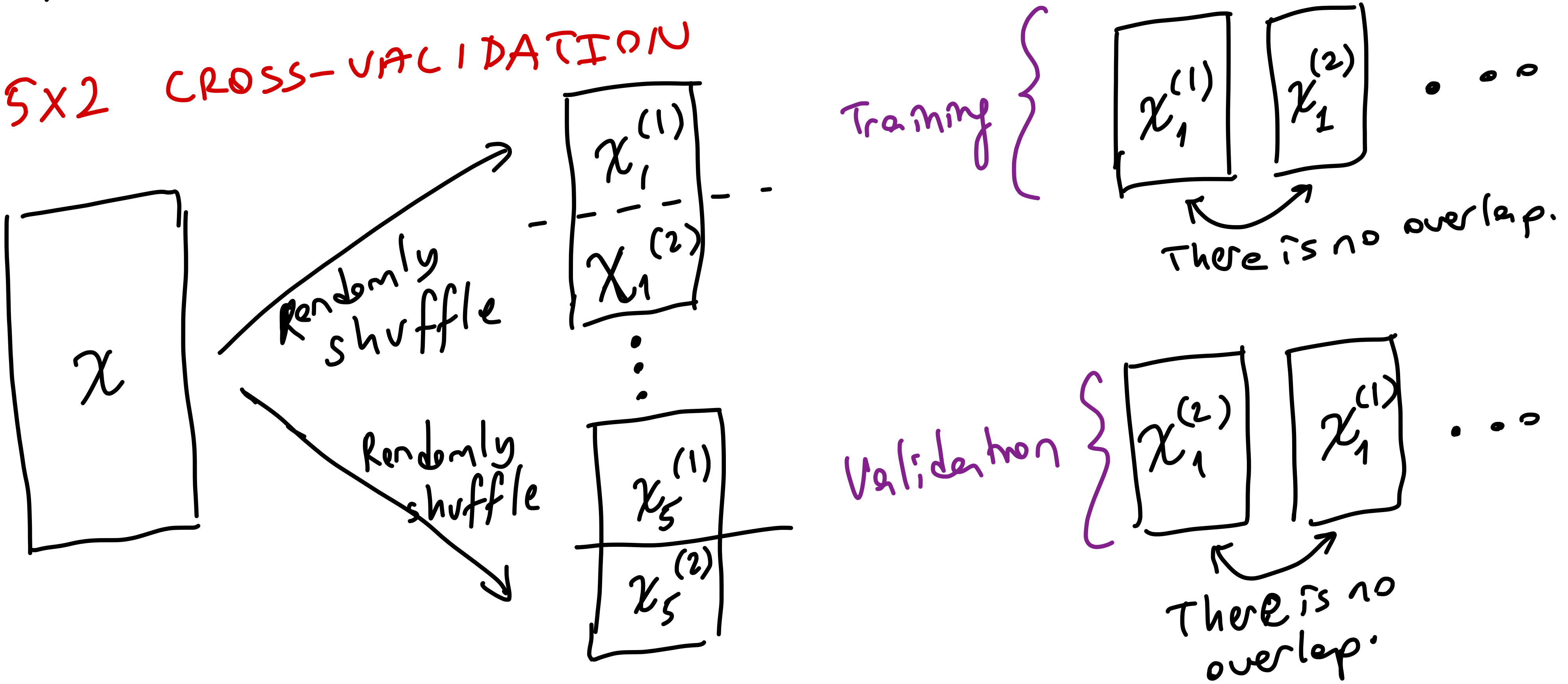
(B) LEAVE-ONE-OUT (LOO) CROSS-VALIDATION

⇒ 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

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

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

Binary Classification

$$\mathcal{X}_{train} = \{(x_i, y_i)\}_{i=1}^{N}$$

$x_i \in \mathbb{R}^D$, $y_i \in \{-1, +1\}$

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

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

$$\mathcal{X}_{test} = \{x_i\}_{i=1}^{N} \qquad \{y_i\}_{i=1}^{N} \Rightarrow \text{unknown}$$

Confusion matrix

Truth \ predicted	+	−	
+	tp	fn	P
−	fp	tn	n
	p'	n'	N

P → # of positives

n → # of negatives

N → dataset size

p' → # of positive predictions

n' → # of negative predictions

2x2

tp = true positive.

tn = true negative

fp = false positive

fn = false negative.

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

$tp + tn \rightarrow$ np.sum(np.diag(.))

$tp + fn + fp + tn \rightarrow$ np.sum (.)

$$\text{Misclassification Error} = \frac{fp + fn}{\underbrace{tp + fn + fp + tn}_{N}} = 1 - \text{Accuracy}.$$

$$\text{tp rate} = tp/p$$
$$\text{fp rate} = fp/n$$
$$\text{precision} = tp/p'$$

$$\text{recall} = tp/p$$
$$\text{sensitivity} = tp/p$$
$$\text{specificity} = tn/n$$

Imbalanced dataset $\left.\begin{bmatrix} 98\% & + \\ 2\% & - \end{bmatrix}\right\}$ predicts positive (+) all the time.

truth \ pred	+	−
+	98	0
−	2	0

95	3
0	2

Accuracy $= \frac{98}{100} = 0.98$ $\quad \underline{\underline{0.97}}$

Precision $= 98/100 = 0.98$ $\quad 95/95 = 1.00$

Recall/Sens. $= 98/98 = 1.00$ $\quad 95/98$

Specificity $= 0/2 = 0.00$ $\quad 2/2 = 1.00$

RECEIVER OPERATING CHARACTERISTICS (ROC) CURVE

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

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

Step #2: Look at their correct labels.

Step #3: Calculate performance statistic with different thresholds.

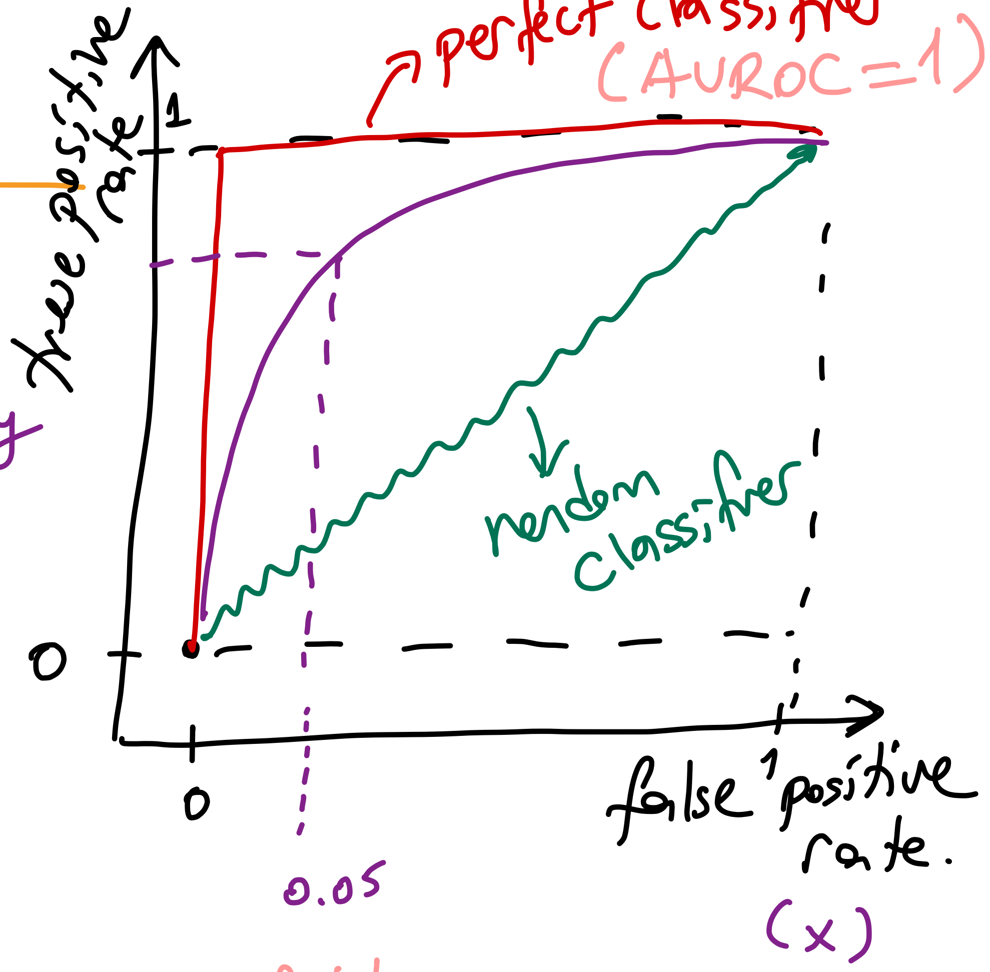
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
	x_8	x_7	x_6	x_1	x_4	x_2	x_3	x_9	x_5
Truth	−	+	−	−	+	+	−	+	+
Ideal	−	−	−	−	+	+	+	+	+
Threshold = 0.45 (prediction)	−	− ⊗	−	−	+	+	+ ⊗	+	+
Threshold = 0.65 (prediction)	−	− ⊗	−	−	− ⊗	− ⊗	+ ⊗	+	+

T \ P	+	−
+	4	1
−	1	3

T \ P	+	−
+	2	3
−	1	3

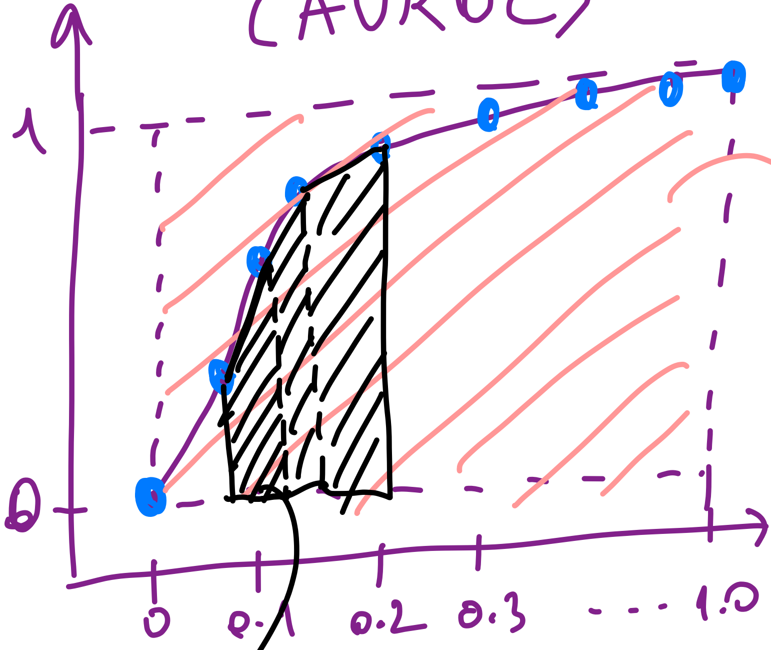
ROC CURVE

$f(x) = y$



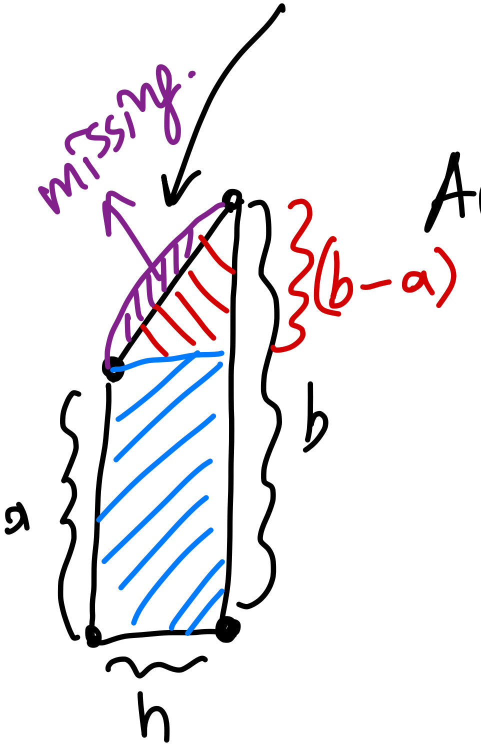
$$\int_0^1 f(x)\,dx \Rightarrow \text{Area under the ROC curve.}$$

(AUROC)



Area of the square is 1.

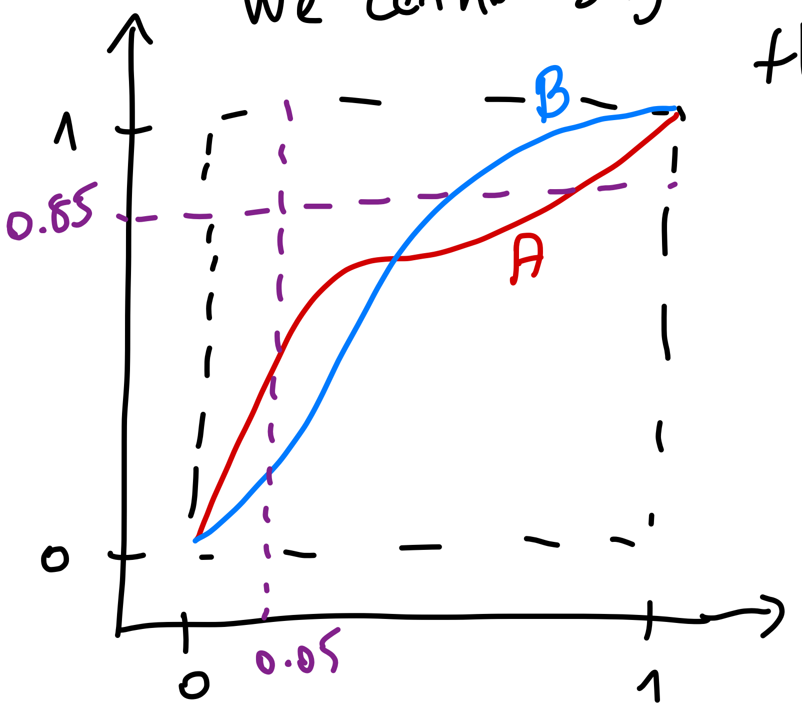
"trapezoid"



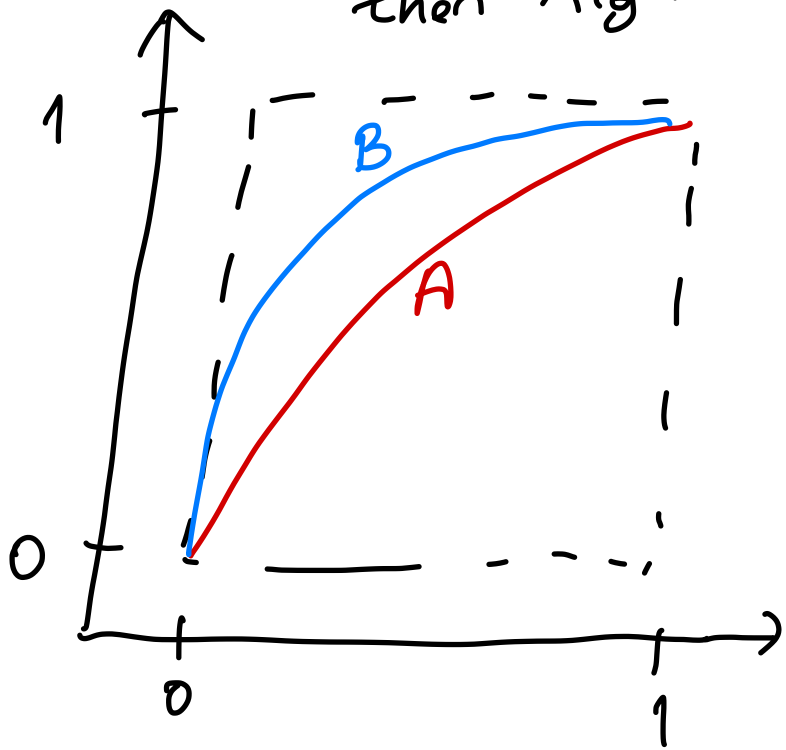
Area of the trapezoid is

$$\frac{(a+b).h}{2} = a.h + \frac{(b-a).h}{2}$$
$$= \frac{(a+b).h}{2}$$

We cannot say one is better then the other.



Alg B is always better then Alg A.



P(+)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Truth (y_i)	−	+	−	−	+	+	−	+	+
Ideal	−	−	−	−	+	+	+	+	+

$$\text{AUROC} = \frac{\#\text{ of correctly placed } (-,+) \text{ pairs}}{\#\text{ of possible } (-,+) \text{ pairs}} = \frac{5+4+4+2}{20} = \frac{15}{20} = \underline{\underline{0.75}}$$

"Wilcoxon-Mann-Whitney statistic"

$\Downarrow$

n.p

P(+) for negatively labeled data points should be smaller then P(+) for positively labeled data points.