# CROSS-VALIDATION METHODS

## K-FOLD CROSS VALIDATION:

| $x_1$ | $x_2$ | . . . | $x_K$ |
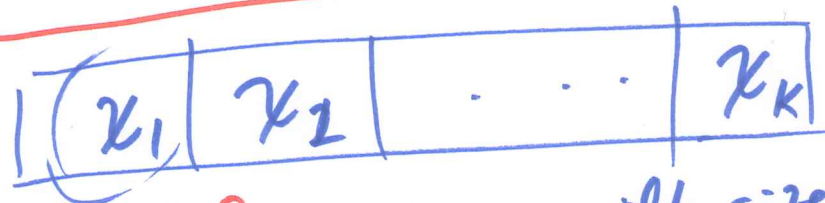
$c_1$ % 40
$c_2$ % 30
$c_3$ % 30

K equally sized partitions

!! Stratification

**Training.**

$1^{st}$ | $x_2, x_3, \ldots, x_k$ |

$2^{nd}$ | $x_1, x_3, \ldots, x_k$ |

:
:

$K = \textcircled{N} \Rightarrow$ LEAVE ONE OUT CV
↳ very small $\Rightarrow$ biomedical applications.

**Validation**

| $x_1$ | $\rightarrow e_{1,1} \quad e_{1,2}$

| $x_2$ | $\rightarrow e_{2,1} \quad e_{2,2}$

:

$\dfrac{e_{k,1}}{\overline{e}_{,1}} \quad \dfrac{e_{k,2}}{\overline{e}_{,2}}$

← expected error rate.

## 5x2 CROSS-VALIDATION



$x \rightarrow$ | $z_1^{(1)}$ | $x_1^{(2)}$ |

| $x_2^{(1)}$ | $x_2^{(2)}$ |

| . | . |

| . | . |

| . | . |
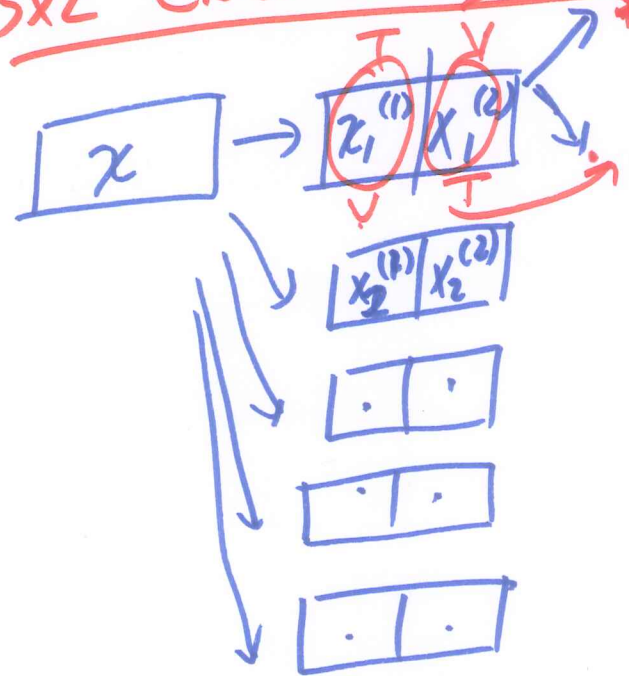
$5 \times 2 = 10$ replications

# MEASURING CLASSIFIER PERFORMANCE

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

Binary Classification $\Rightarrow$ $\chi_{train} = \{(x_i, y_i)\}_{i=1}^{N_{train}}$

$x_i \in \mathbb{R}^P$

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

$f : x \rightarrow y$

$\chi_{test} = \{(x_i)\}_{i=1}^{N_{test}}$

$\boxed{y_i = ?}$

Predicted

|  | + | − |  |
|---|---|---|---|
| + | TP | FN | P |
| − | FP | TN | n |
|  | P' | n' |  |

Truth

$f(x_1) \Rightarrow \hat{y_1}$
$f(x_2) = \hat{y_2}$
$\vdots$
$f(x_{N_{test}}) = \hat{y}_{TEST}$

? measure performance.

$$TP + FP + FN + TN = N_{test}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad N_{test}$$

$$\text{Accuracy} + \text{Misclass. Error} = 1$$

$$\text{Missclassification Error} = \frac{FP + FN}{N_{test}}$$

②

TP rate $\Rightarrow$ TP/P

FP rate $\Rightarrow$ FP/(n)

precision $\Rightarrow$ TP/P'

recall $\Rightarrow$ TP/P

sensitivity $\Rightarrow$ TP/P

specificity $\Rightarrow$ TN/(n)

$1-(FP\text{-rate}) = $ specificity.

$1-$ specificity $= (FP\text{-rate})$

# RECEIVER OPERATING CHARACTERISTICS (ROC) CURVE

$f \Rightarrow$ returns
$+$
all the true

98% +
2% −
___
98% − it will be 98%
2% + predicted accurate.

AUROC

Area under the ROC curve

AUC
Area under the curve.

AUPR

|   | + | − |
|---|---|---|
| + | 98 | 0 |
| − | 2 | 0 |

FP

|   |   | $\to$ FN |
|---|---|---|
|   | 0 | 2 |
|   | 0 | 98 |

$P(+) \Rightarrow 0.4, 0.6, 0.7, 0.5, 0.9, 0.3, 0.2,$
$0.1, 0.8 \to \theta = 0.35$

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

(3)

TP-rate

random classifier

Area = 1.
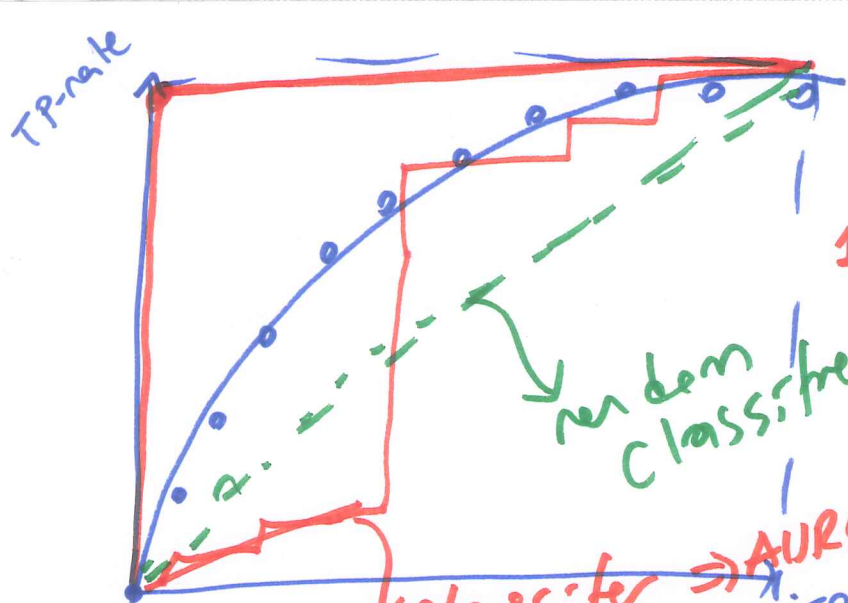
classifier

$\rightarrow$ AUROC $> 0.5$
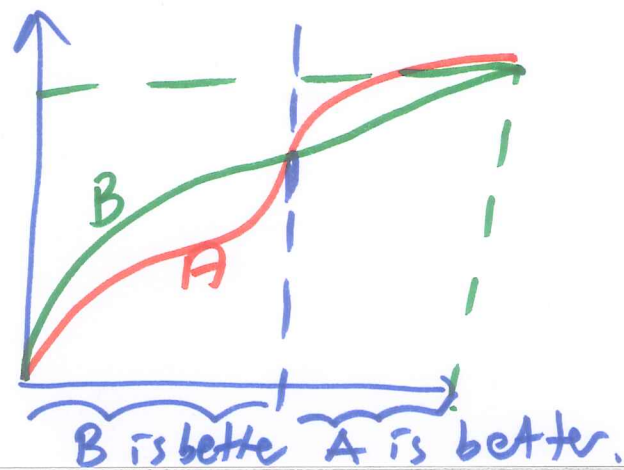
FP-rate
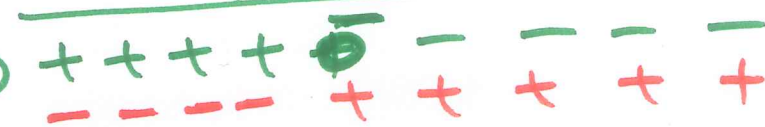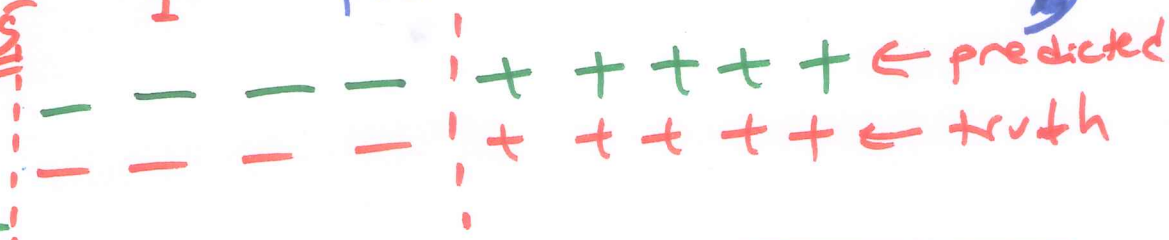
perfect classifier $\Rightarrow$ AUROC $= \underline{1}$

perfectly incorrect classifier
$\Rightarrow$ AUROC $= 0$

random classifier $\Rightarrow$ AUROC $\tilde{=} 0.5$

|   | + | − |
|---|---|---|
| + | 4 | 1 |
| − | 2 | 2 |

Truth

predicted

$FP\text{-rate} = \dfrac{2}{4}$

$TP\text{-rate} = \dfrac{4}{5}$

$- - - - - | + + + + + \leftarrow$ predicted

$- - - - - | + + + + + \leftarrow$ truth

$+ + + + \overline{\delta} - - - - -$

$- - - - + + + + +$

1

B

A
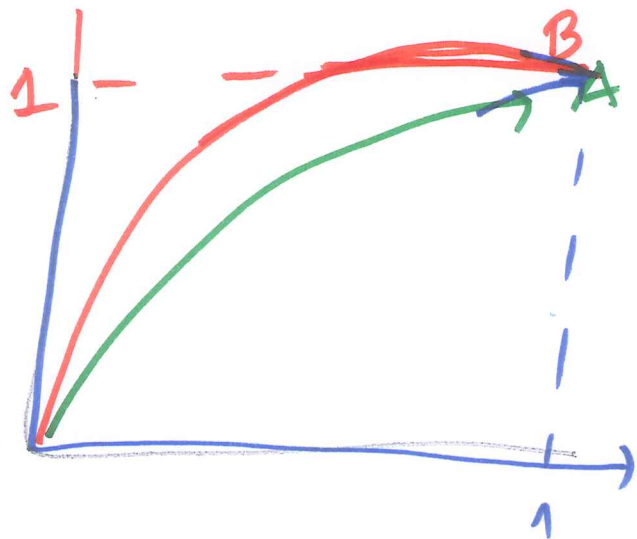
B is better than A because it is always above A.

B

A

B is better   A is better.

$$P(+) \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4 \quad 0.5 \quad 0.6 \quad 0.7 \quad 0.8 \quad 0.9$$

$$\boxed{-}\ \ \pm\ \ =\ \ =\ \ \pm\ \ \pm\ \ =\ \ \pm\ \ \pm$$

# of negatives = 4
# of positives = 5

$$\frac{15}{20} = \frac{5 + 4 + 4 + 2}{20} = 0.75$$

$$AUROC = \frac{\#\ of\ correctly\ placed\ (-, +)\ pairs}{\#\ of\ possible\ (-, +)\ pairs}$$

**Multiclass classification:**



$$Accuracy = \frac{summation\ of\ diagonal\ entries}{summation\ of\ all\ entries.}$$

# Information Retrieval
— retrieve relevant documents

retrieved documents ⟶ ← relevant documents.

b | a | c

misses ⟵ hits
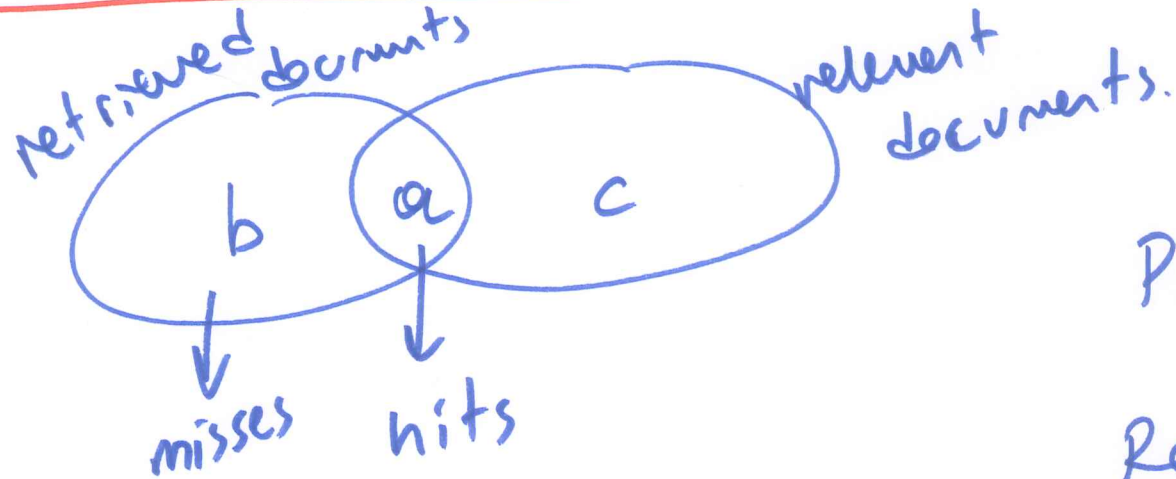
$$Precision = \frac{a}{a+b}$$

$$Recall = \frac{a}{a+c}$$

ROC ⟹ PR curves