Machine Learning

Concepts et méthodologie

marie.szafranski@ensiie.fr

What's now ?

Méthodologie

1. Comment évaluer les performances d'un modèle ?
   - Mesures de performance « brutes »
     - ⇝ Matrice de confusion
     - ⇝ Courbe ROC
   - Procédures d'estimation du risque empirique
     - ⇝ Découpage
     - ⇝ Simulation
     - ⇝ Pénalisation

(Vignette Wikistat)

http://wikistat.fr/pdf/st-m-app-risque.pdf

2. Comment choisir un bon modèle ?
   - ⇝ Inclusion dans les procédures d'estimation du risque empirique

Contexte : la classification binaire

Minimisation du risque empirique

⇝ Cf. 01_concepts.pdf

- Ensemble d'apprentissage

$$S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$$
$$\forall i,\ \boldsymbol{x}_i \in \mathcal{X} = \mathbb{R}^M$$
$$\forall i,\ y_i \in \mathcal{Y} = \{0, 1\} \text{ ou } \{\pm 1\}$$

- Construire un modèle

$$\hat{y} = \text{sign}\left[h^\star_{S,p}(\boldsymbol{x})\right]$$
$$h^\star_{S,p} : \mathbb{R}^M \to \mathcal{Y}$$

- Erreur empirique : estimation optimiste de l'erreur de généralisation

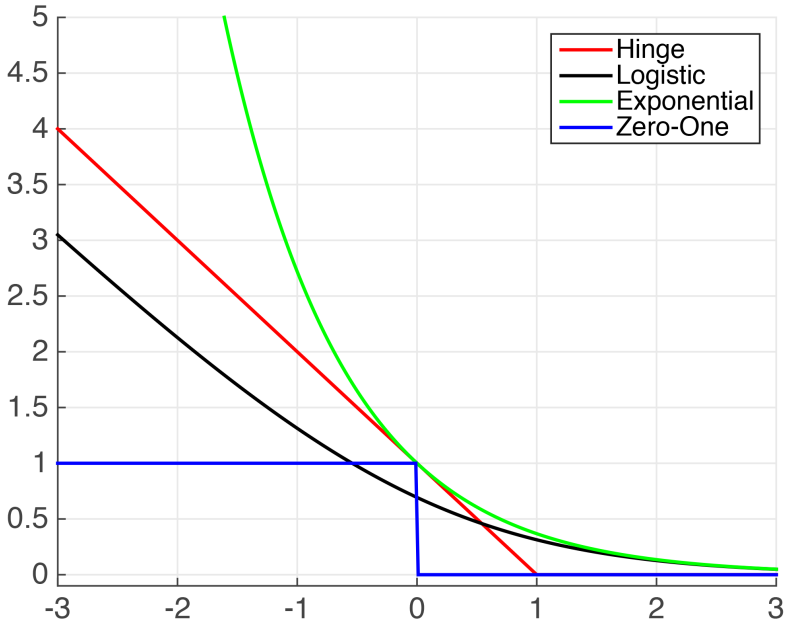
$$\begin{aligned} h^\star_{S,p} &= \text{argmin}_{h \in \mathcal{H}} \quad R_{emp}(h) \\ &= \text{argmin}_{h \in \mathcal{H}} \quad \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(\boldsymbol{x}_i)) \end{aligned}$$

Contexte : la classification binaire

⇝ Cf. 01_concepts.pdf

Fonction de perte

- Pénalisation de l'erreur
  ± sévère
- Mesure de performance ?



Crédits : http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote10.html

Contexte : la classification binaire

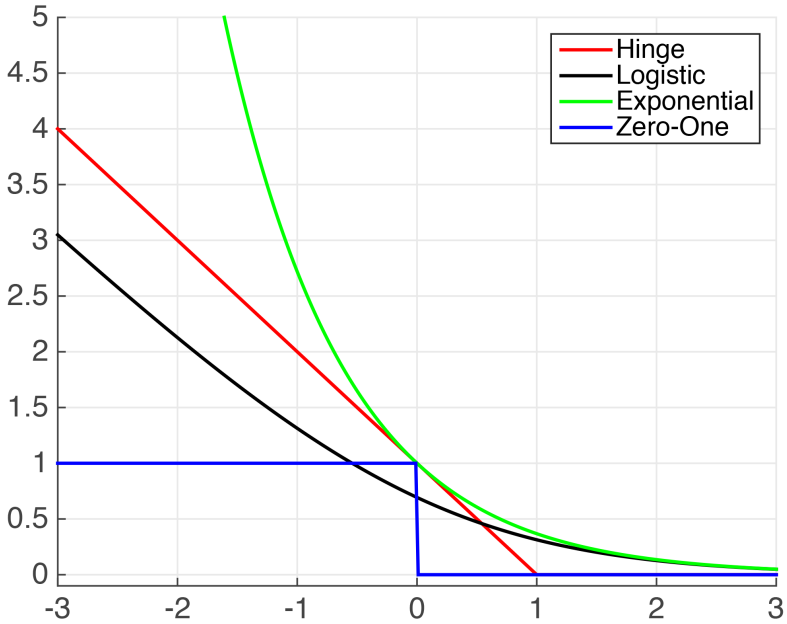
⇝ Cf. 01_concepts.pdf

Fonction de perte

- Pénalisation de l'erreur
  ± sévère

- Mesure de performance ?

$$\hat{y} = \text{sign}\left[h^{\star}_{S,p}(\boldsymbol{x})\right]$$



Crédits : http://www.cs.cornell.edu/courses/cs4780/2015fa/web/lecturenotes/lecturenote10.html

Contexte : la classification binaire

Mesurer l'erreur en pratique

Supervisé

- Matrice de confusion

  Mesures associées $\rightsquigarrow$ Taux d'erreur, précision, rappel, ...

- Courbe ROC

  Mesures associées $\rightsquigarrow$ Sensibilité, spécificité, aire sous la courbe, ...

Contexte : la classification binaire

Mesurer l'erreur en pratique

Supervisé

- Matrice de confusion

  Mesures associées $\rightsquigarrow$ Taux d'erreur, précision, rappel, ...

- Courbe ROC

  Mesures associées $\rightsquigarrow$ Sensibilité, spécificité, aire sous la courbe, ...

Classification ($C > 2$) ou régression

$\Rightarrow$ Adaptations nécessaires

Matrice de confusion

Terminologie

Observations → / Prédictions ↓	$y = 1$	$y = 0$	Total
$\hat{y} = 1$	$N_{\hat{1}1}$	$N_{\hat{1}0}$	$N_{\hat{1}.}$
$\hat{y} = 0$	$N_{\hat{0}1}$	$N_{\hat{0}0}$	$N_{\hat{0}.}$
Total	$N_{.1}$	$N_{.0}$	N

- # observations positives ($N_{.1}$, Pos) et négatives ($N_{.0}$, Neg)
- # prédictions positives ($N_{.\hat{1}}$, $\widehat{\text{Pos}}$) et négatives ($N_{.\hat{0}}$, $\widehat{\text{Neg}}$)

- # vrais positifs ($N_{\hat{1}1}$, VP, TP) et négatifs ($N_{\hat{0}0}$, VN, TN) — #préd. correctes
- # faux positifs ($N_{\hat{1}0}$, FP) et négatifs ($N_{\hat{0}1}$, FN) — #préd. erronées

Matrice de confusion

Mesure de performance globale

Observations → / Prédictions ↓	$y = 1$	$y = 0$	Total
$\hat{y} = 1$	TP	FP	$\widehat{\text{Pos}}$
$\hat{y} = 0$	FN	TN	$\widehat{\text{Neg}}$
Total	Pos	Neg	N

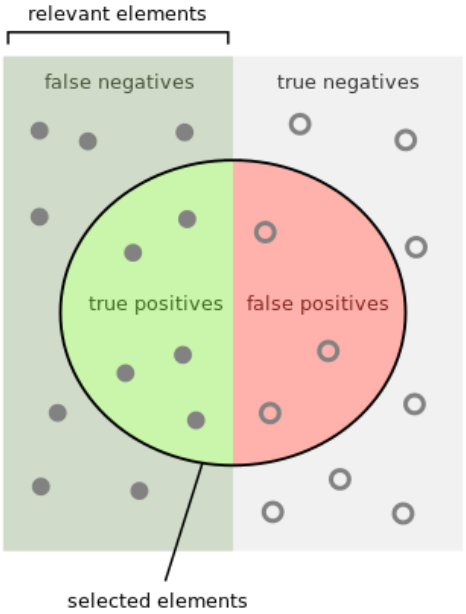
$$\frac{\text{TP} + \text{TN}}{N} = 1 - \frac{\text{FP} + \text{FN}}{N}$$

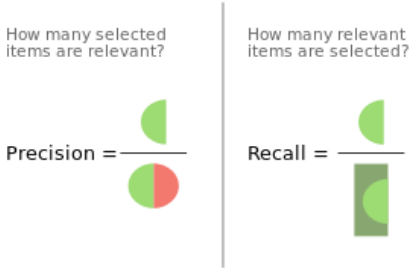
$$\text{Précision globale} = 1 - \text{taux d'erreur}$$

Accuracy vs Error rate

Matrice de confusion

Précision





Crédits : Wikipedia

Observations → / Prédictions ↓	$y = 1$	$y = 0$	Total
$\hat{y} = 1$	TP	FP	$\widehat{\text{Pos}}$
$\hat{y} = 0$	FN	TN	$\widehat{\text{Neg}}$
Total	Pos	Neg	N

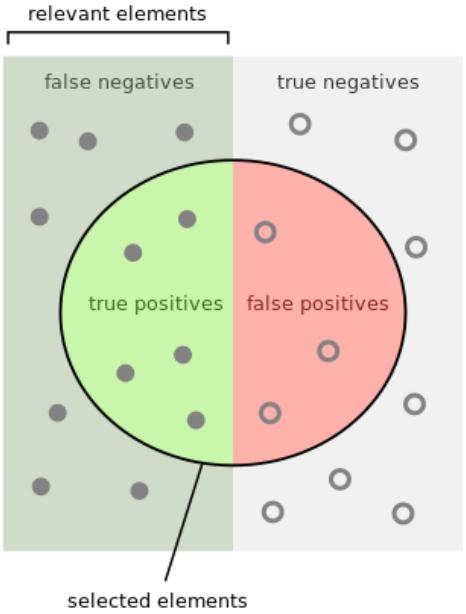
Combien d'éléments sélectionnés sont pertinents ?
(parmi tous les éléments sélectionnés : $\hat{y} = 1$)

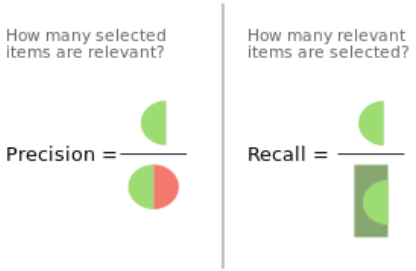
$$\frac{\text{TP}}{\widehat{\text{Pos}}} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

Precision, Positive Predictive Value (PPV)

Matrice de confusion

Rappel





Crédits : Wikipedia

Taux de vrais positifs, sensitivité

Observations → / Prédictions ↓	$y = 1$	$y = 0$	Total
$\hat{y} = 1$	TP	FP	$\widehat{\text{Pos}}$
$\hat{y} = 0$	FN	TN	$\widehat{\text{Neg}}$
Total	Pos	Neg	N

Combien d'éléments pertinents sont sélectionnés ?
(parmi tous les éléments pertinents : $y = 1$)

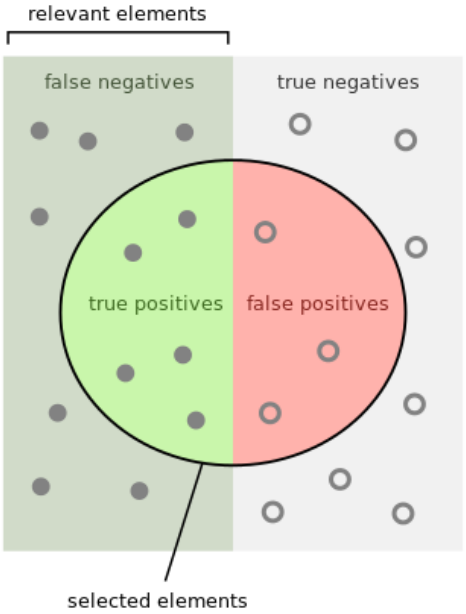
$$\frac{\text{TP}}{\text{Pos}} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

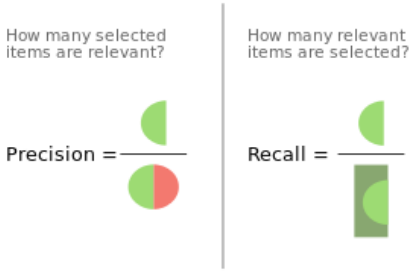
Recall, True Positive Rate (TPR), sensitivity

Matrice de confusion

Spécificité

Taux de vrais négatifs, sélectivité





Crédits : Wikipedia

Observations → / Prédictions ↓	$y = 1$	$y = 0$	Total
$\hat{y} = 1$	TP	FP	$\widehat{\text{Pos}}$
$\hat{y} = 0$	FN	TN	$\widehat{\text{Neg}}$
Total	Pos	Neg	N

Combien d'éléments pertinents sont sélectionnés ?
(parmi tous les éléments pertinents : $y = 0$)

$$\frac{\text{TN}}{\text{Neg}} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

True Negative Rate (TNR), selectivity

Matrice de confusion

Autres taux d'erreurs et mesures dérivées

- Taux de faux positifs — False Positive Rate (FPR)

$$\frac{\mathsf{FP}}{\mathsf{Neg}} = \frac{\mathsf{FP}}{\mathsf{TN}+\mathsf{FP}} = 1 - \frac{\mathsf{TN}}{\mathsf{Neg}} = 1 - \text{spécificité}$$

- Taux de faux négatifs — False Negative Rate (FNR)

$$\frac{\mathsf{FN}}{\mathsf{Pos}} = \frac{\mathsf{FN}}{\mathsf{TP}+\mathsf{FN}} = 1 - \frac{\mathsf{TP}}{\mathsf{Pos}} = 1 - \text{rappel}$$

- Taux de fausses découvertes — False Discovery Rate (FDR)

$$\frac{\mathsf{FP}}{\widehat{\mathsf{Pos}}} = \frac{\mathsf{FP}}{\mathsf{TP}+\mathsf{FP}} = 1 - \frac{\mathsf{TP}}{\widehat{\mathsf{Pos}}} = 1 - \text{précision}$$

- F1-score

$$F_1 = 2 \times \frac{\mathsf{PPV} \times \mathsf{TPR}}{\mathsf{PPV}+\mathsf{TPR}} = 2 \times \frac{\mathsf{TP}}{(\mathsf{TP}+\mathsf{FP}+\mathsf{FN})}$$

Matrice de confusion

Récapitulatif et mesures supplémentaires

- Vignette Wikistat ⇝ récap. + point de vue complémentaire
  `http://wikistat.fr/pdf/st-m-app-risque.pdf`

- Article Wikipedia ⇝ mesures complémentaires
  `https://en.wikipedia.org/wiki/Precision_and_recall`

Courbe ROC — Receiver Operating Caracteristic

Définition informelle et intuitions

Évaluation au *seuil* s d'un algorithme retournant un *score*

$\rightsquigarrow$ probabilité qu'un exemple soit positif

- *Score* $h_{sc}(\boldsymbol{x}) \in [0,1]$

  Artificiel K-nn : $h_{sc}(\boldsymbol{x}) = \frac{\#_K(\boldsymbol{x}|y=1)}{K}$

  Intrinsèque Régression logistique : $h(\boldsymbol{x}) = \frac{1}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta})}$

- *Seuil*

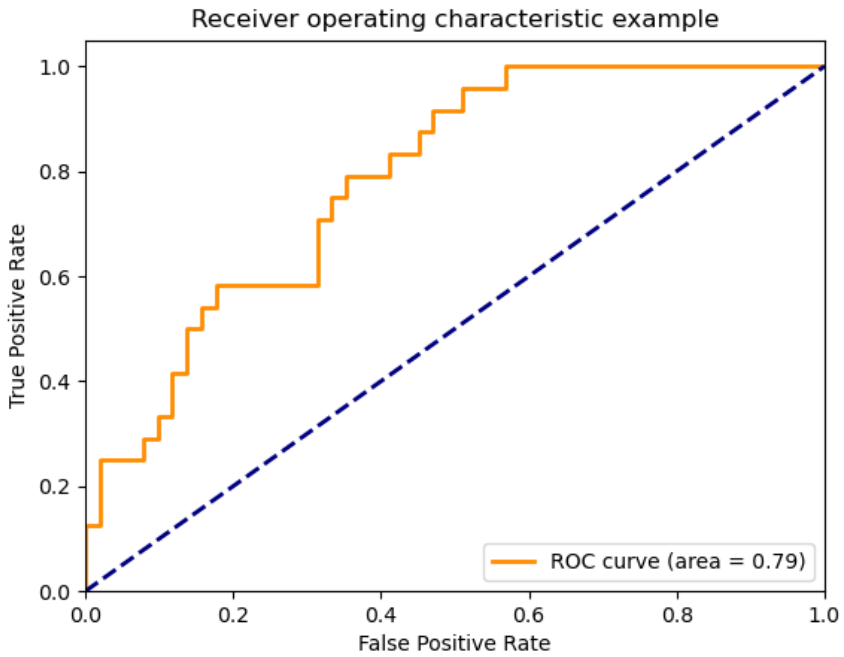
$$\hat{y} = 1 \quad \text{si } h(\boldsymbol{x}) \geq s$$
$$\hat{y} = 0 \quad \text{si } h(\boldsymbol{x}) < s$$

Courbe ROC

Représentation graphique

↑ proba de détecter un vrai signal (TPR = sensibilité)

→ proba de détecter un signal à tord (FPR = 1 - spécificité)



Crédits : `scikit-learn`

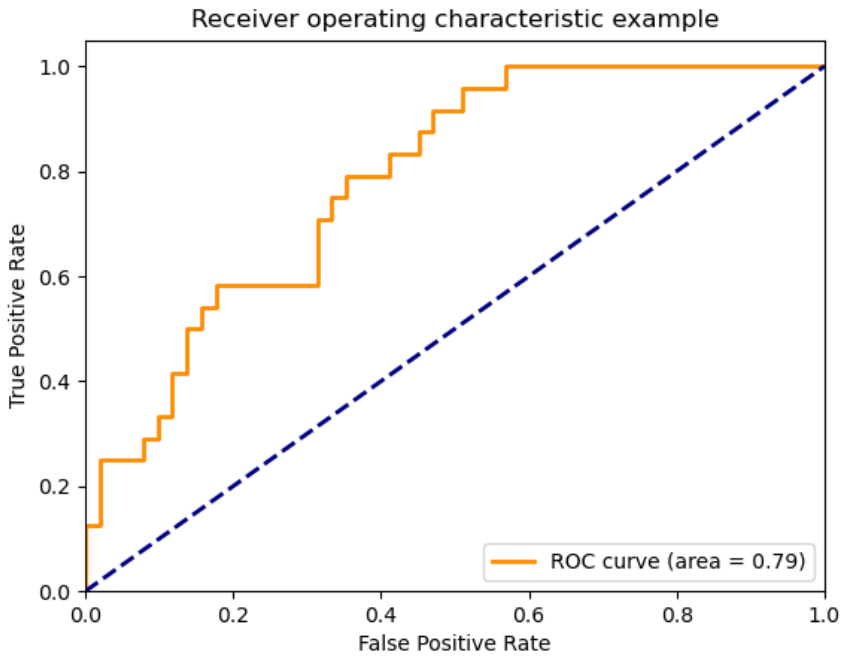
- Point de la courbe *Seuil* s

Courbe ROC

Représentation graphique

↑ proba de détecter un vrai signal (TPR = sensibilité)

→ proba de détecter un signal à tord (FPR = 1 - spécificité)



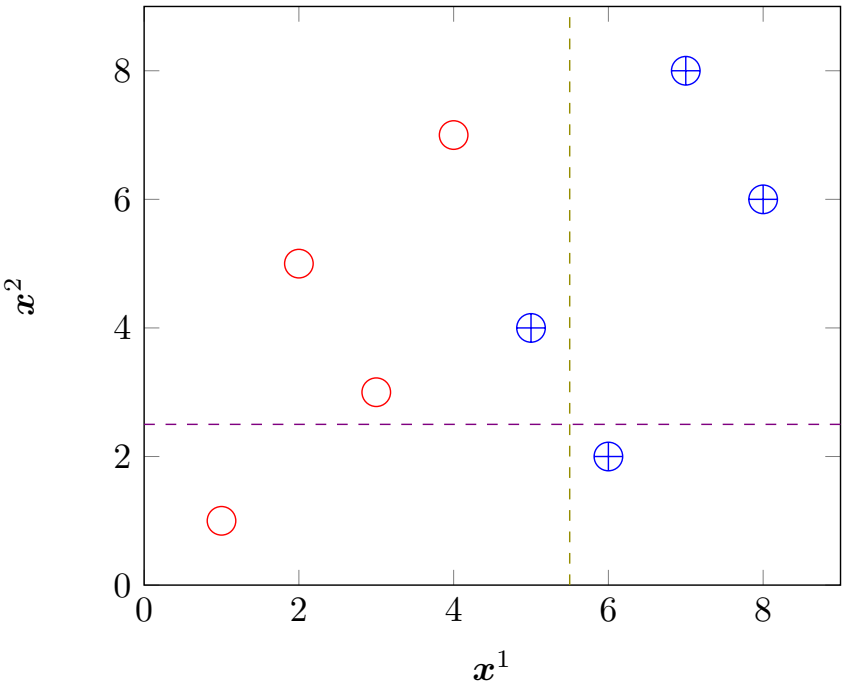
Crédits : `scikit-learn`

- Point de la courbe — *Seuil* s
- Area Under the Curve — AUC

Courbe ROC

Illustration

Exemple tiré du cours de M. Mougeot



◯ $y = 0$ et ⊕ $y = 1$

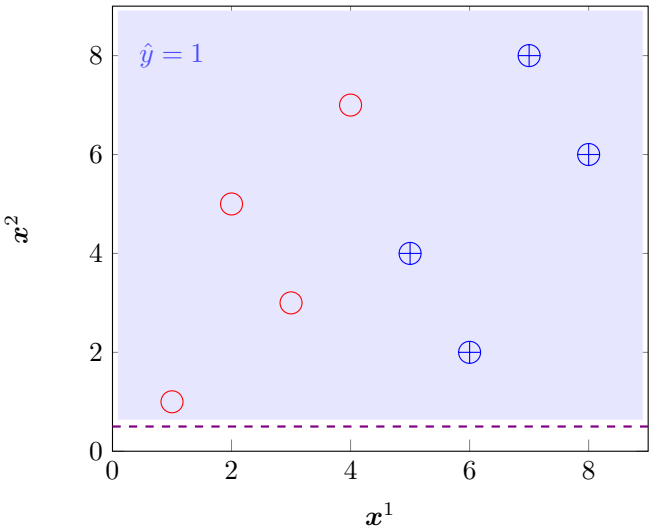
1. $h(\boldsymbol{x}) = \boldsymbol{x}^1 \geq s \in [0, 9]$

2. $h(\boldsymbol{x}) = \boldsymbol{x}^2 \geq s \in [0, 9]$

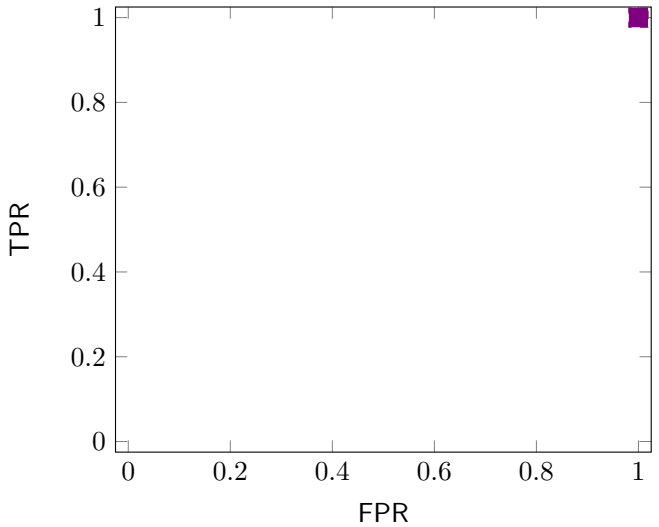
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^2$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{4}{4 + 0} = 1.00 \qquad \text{sensitivité}$$

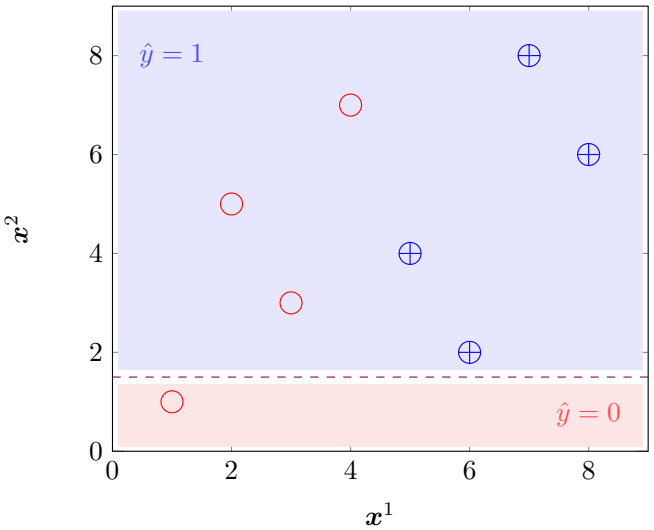
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{0}{0 + 4} = 0.00 \qquad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \qquad 1 - \text{spécificité}$$

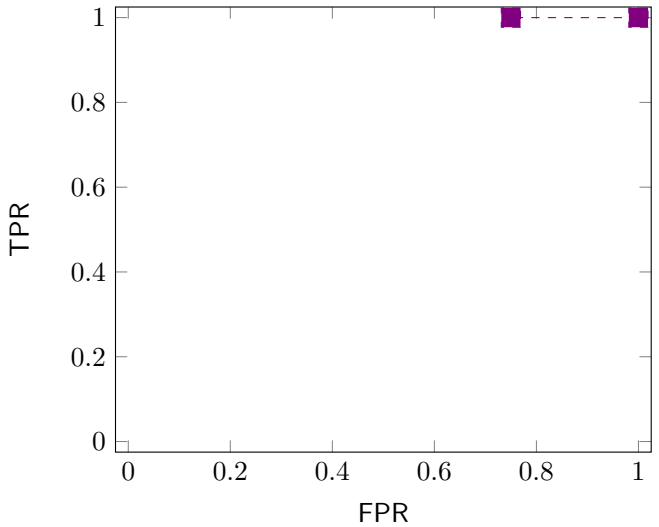
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^2$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00
1.50	1.00	0.25	0.75



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{4}{4+0} = 1.00 \qquad \text{sensitivité}$$

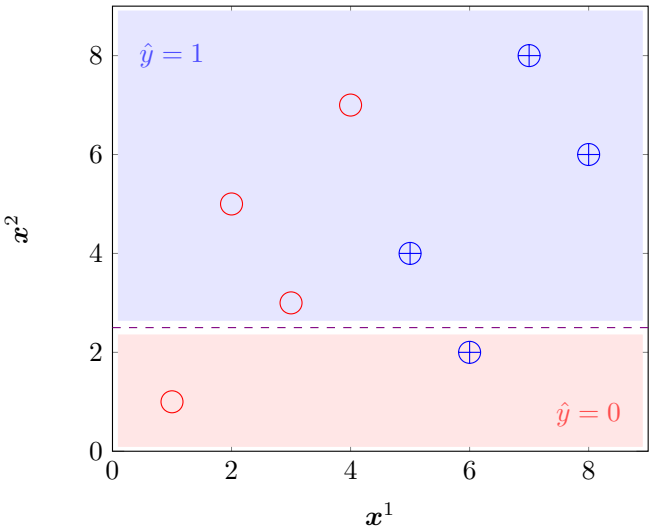
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{1}{1+3} = 0.25 \qquad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \qquad 1 - \text{spécificité}$$

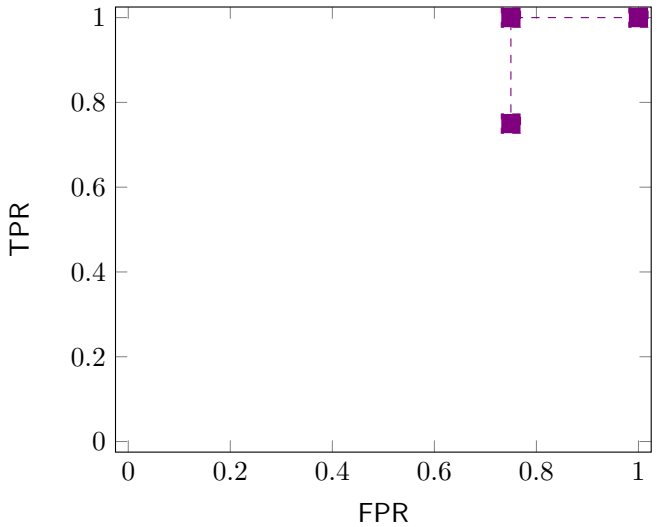
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^2$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00
1.50	1.00	0.25	0.75
2.50	0.75	0.25	0.75



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3}{3+1} = 0.75 \qquad \text{sensitivité}$$

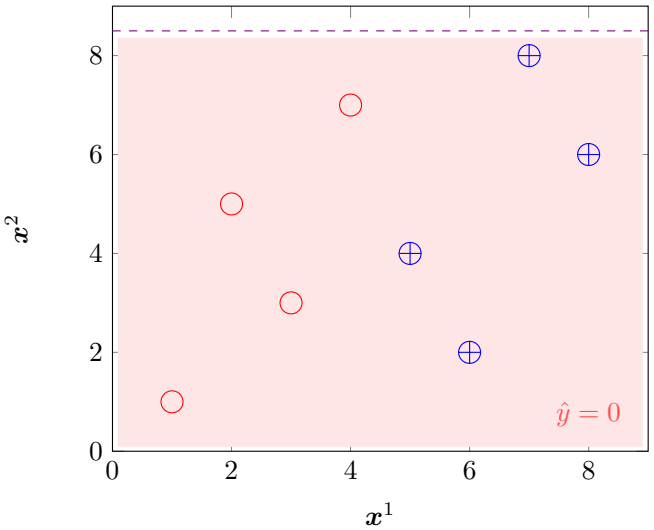
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{1}{1+3} = 0.25 \qquad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \qquad 1 - \text{spécificité}$$

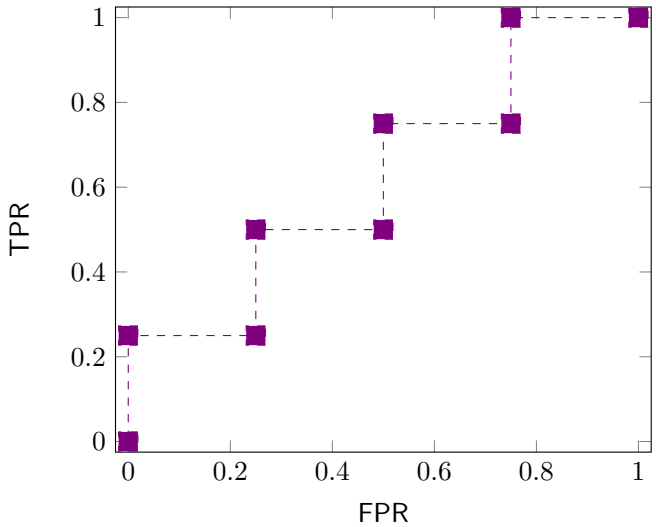
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^2$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00
1.50	1.00	0.25	0.75
2.50	0.75	0.25	0.75
3.50	0.75	0.50	0.50
4.50	0.50	0.50	0.50
5.50	0.50	0.75	0.25
6.50	0.25	0.75	0.25
7.50	0.25	1.00	0.00
8.50	0.00	1.00	0.00



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{0}{0 + 4} = 0.00$$ sensitivité

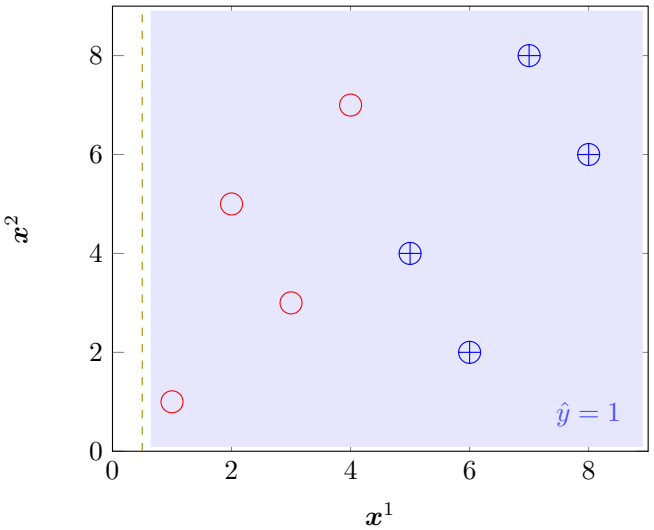
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{4}{4 + 0} = 1.00$$ spécificité

$$\text{FPR} = 1 - \text{TNR}$$ $1 -$ spécificité

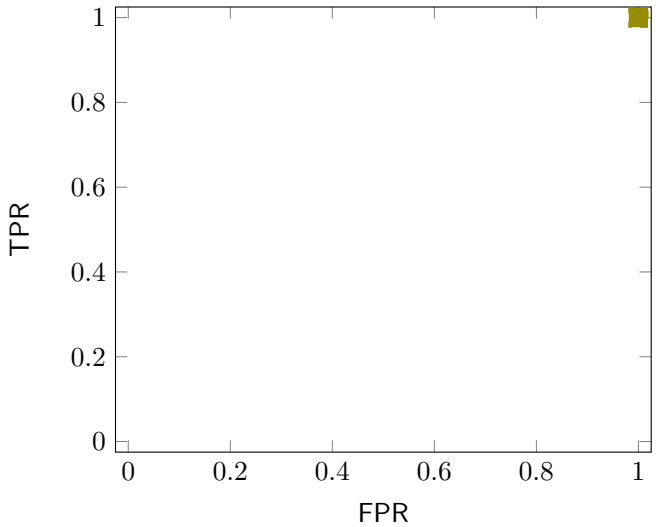
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^1$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{4}{4+0} = 1.00 \qquad \text{sensitivité}$$

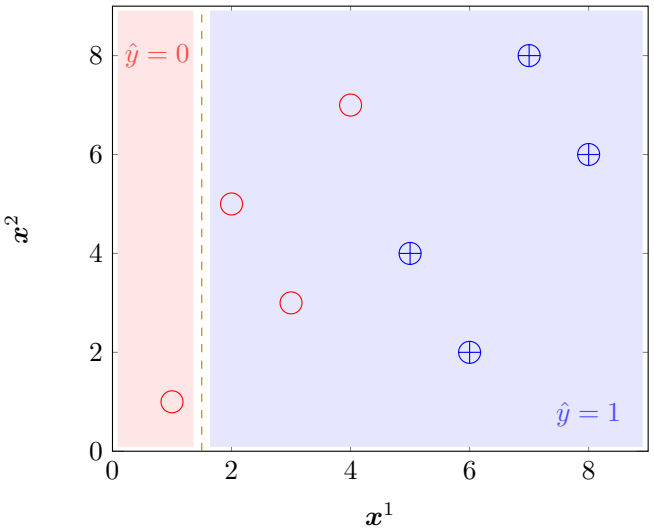
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{0}{0+4} = 0.00 \qquad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \qquad 1 - \text{spécificité}$$

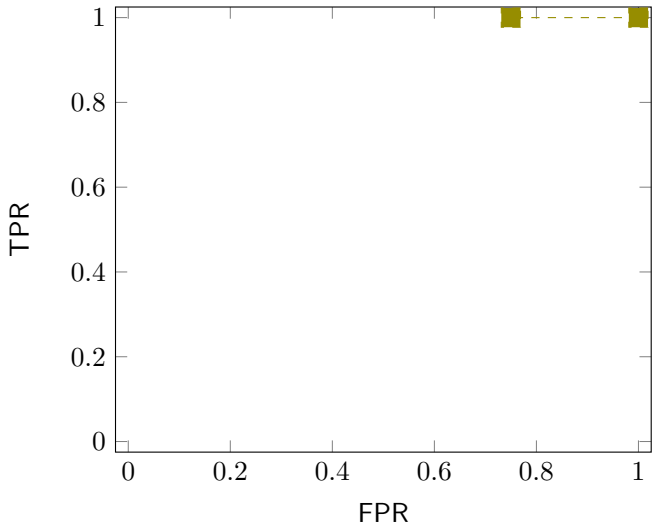
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^1$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00
1.50	1.00	0.25	0.75



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{4}{4+0} = 1.00 \qquad \text{sensitivité}$$

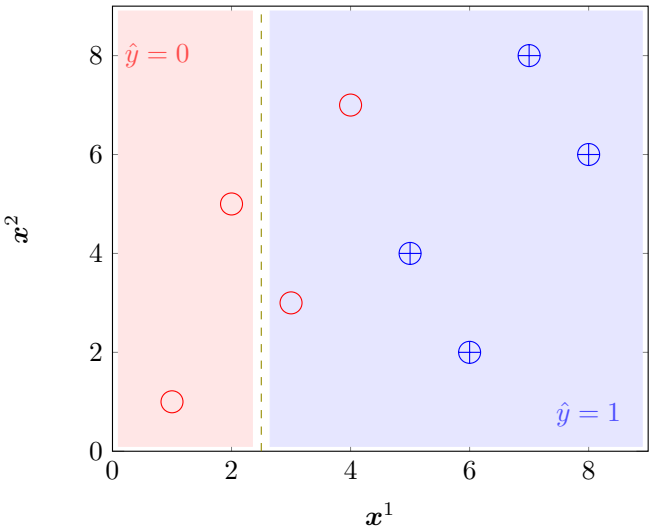
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{1}{1+3} = 0.25 \qquad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \qquad 1 - \text{spécificité}$$

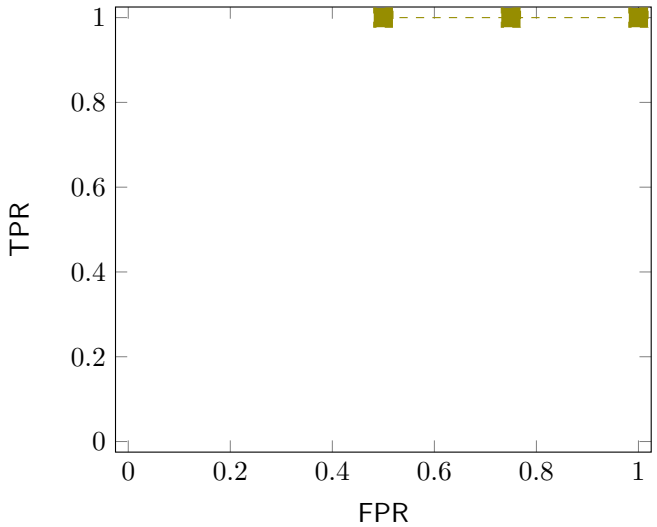
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^1$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00
1.50	1.00	0.25	0.75
2.50	1.00	0.50	0.50



$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{4}{4+0} = 1.00 \qquad \text{sensitivité}$$

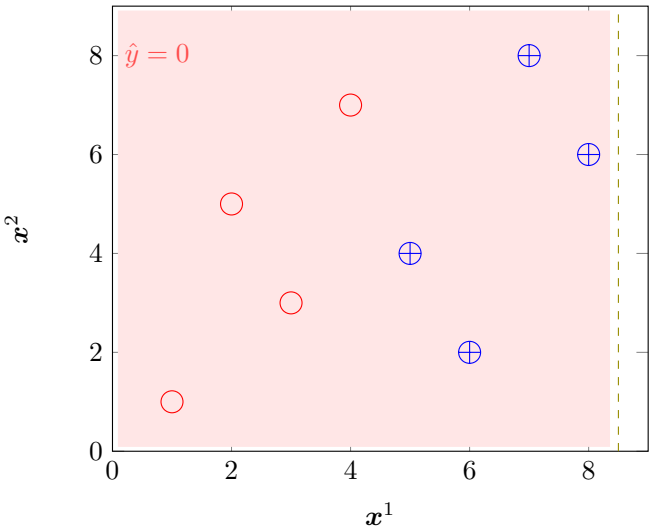
$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{2}{2+2} = 0.50 \qquad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \qquad 1 - \text{spécificité}$$

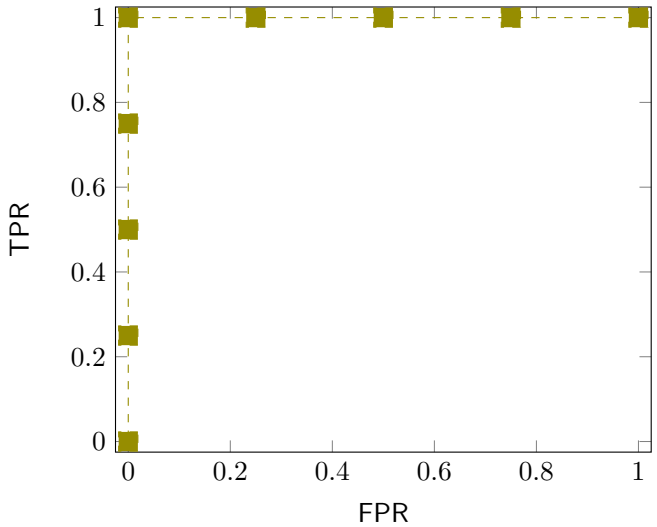
Courbe ROC

Illustration

$h(\boldsymbol{x}) = \boldsymbol{x}^1$



s	TPR	TNR	FPR
0.50	1.00	0.00	1.00
1.50	1.00	0.25	0.75
2.50	1.00	0.50	0.50
3.50	1.00	0.75	0.25
4.50	1.00	1.00	0.00
5.50	0.75	1.00	0.00
6.50	0.50	1.00	0.00
7.50	0.25	1.00	0.00
8.50	0.00	1.00	0.00



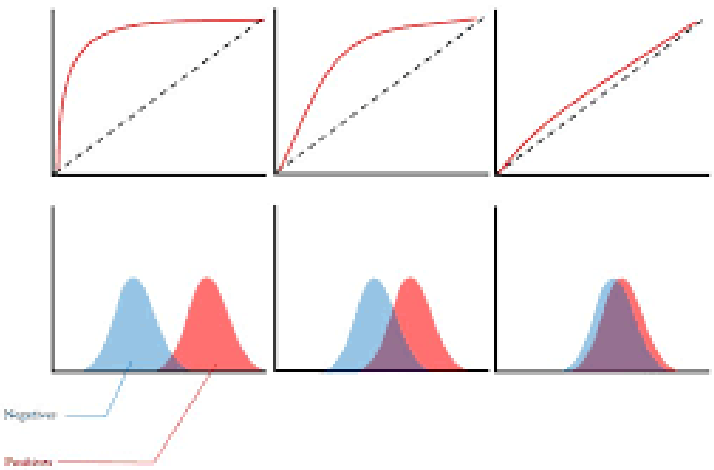
$$\text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{0}{0 + 4} = 0.00 \quad \text{sensitivité}$$

$$\text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}} = \frac{4}{4 + 0} = 1.00 \quad \text{spécificité}$$

$$\text{FPR} = 1 - \text{TNR} \quad 1 - \text{spécificité}$$

Courbe ROC

En résumé : comparaison de différents modèles



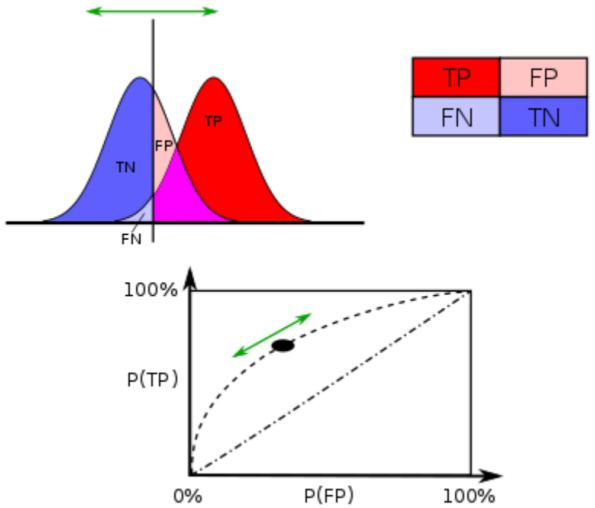
Crédits : Alex Yartsev

- Comparaison globale — AUC

$h_G(\boldsymbol{x})$ à perf. exacte	1.0
$h_M(\boldsymbol{x})$ à perf. "moyenne"	~ 0.75
$h_D(\boldsymbol{x})$ à perf. aléatoire	0.5

Courbe ROC

En résumé : comparaison de différents modèles



Crédits : `en.wikipedia`

- Comparaison locale $h_\ell(\boldsymbol{x}) \geq s$

What else ?

Autres mesures de performance

Plein, vraiment plein...

Cas spécifiques

- Données déséquilibrées
- Multi-classes
- Régression
- Ordonnancement
- ...

⇝ https://scikit-learn.org/stable/modules/model_evaluation.html

Interlude : évaluation des performances avec les K-nn

Algorithme

`https://scikit-learn.org/stable/modules/neighbors.html`

Données Pima

$C = 2$

Évaluation

`https://scikit-learn.org/stable/modules/model_evaluation.html`

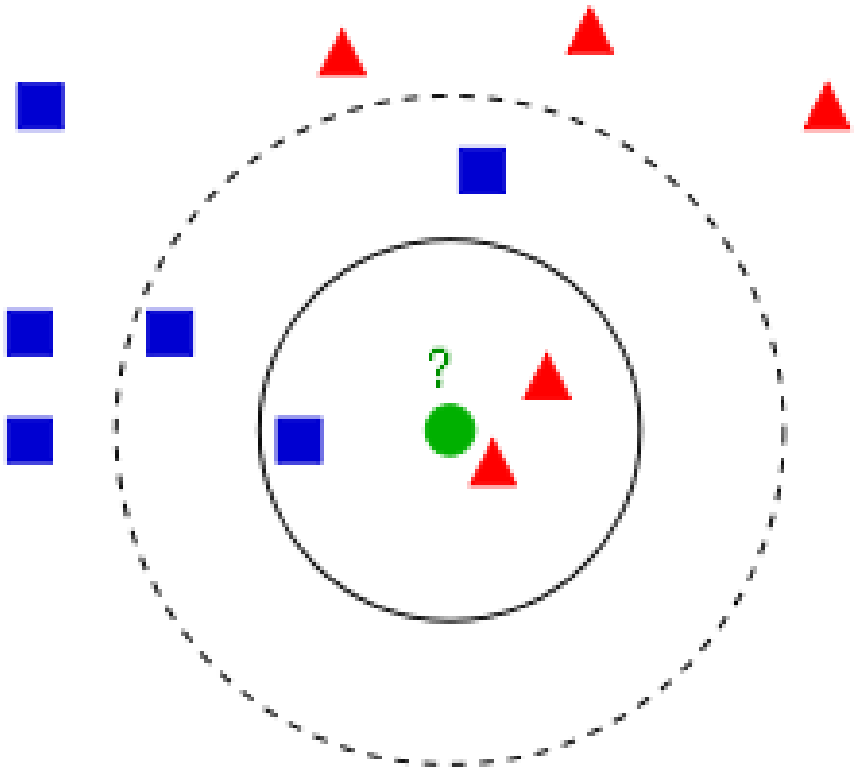
Focus sur les mesures vues aujourd'hui

- $\rightarrow$ Module `classification metrics`
- $\rightsquigarrow$ Extension au cas multiclasse

Les K plus proches voisins (K-nn)

Données $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, $y_i = 1$ ou $y_i = -1$

Modèle h_K avec $K = \{3 —, 5 - - -\}$



Exemple tiré de Wikipedia

Les K plus proches voisins (K-nn)

Entrées $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_\ell}$, $S_u = \{\boldsymbol{x}_j\}_{j=1}^{N_u}$ et K

Sortie $\{\hat{y}_j\}_{j=1}^{N_u}$

Les K plus proches voisins (K-nn)

Entrées $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_\ell}$, $S_u = \{\boldsymbol{x}_j\}_{j=1}^{N_u}$ et K

Sortie $\{\hat{y}_j\}_{j=1}^{N_u}$

1. Calculer les similarités $\mathbf{D} \in \mathbb{R}^{N_u \times N_\ell}$

   $\mathbf{D}_{j,i} = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$ $1 \leq j \leq N_u, 1 \leq i \leq N_\ell$

2. Ordonner les K similarités $\overleftarrow{\mathbf{D}} \in \mathbb{R}^{N_u \times K}$

   $i_1 : \overleftarrow{\mathbf{D}}_{j,i_1} \leq \cdots \leq i_{K-1} : \overleftarrow{\mathbf{D}}_{j,i_{K-1}} \leq i_K : \overleftarrow{\mathbf{D}}_{j,i_K}$

3. Affecter les classes

   $\hat{y}_j = \text{classe\_majoritaire}(y_{i_1}, \cdots, y_{i_K})$ $1 \leq j \leq N_u$

Les K plus proches voisins (K-nn)

Entrées $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_\ell}$, $S_u = \{\boldsymbol{x}_j\}_{j=1}^{N_u}$ et K

Sortie $\{\hat{y}_j\}_{j=1}^{N_u}$

1. Calculer les similarités $\mathbf{D} \in \mathbb{R}^{N_u \times N_\ell}$

   $\mathbf{D}_{j,i} = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$ $1 \leq j \leq N_u, 1 \leq i \leq N_\ell$

2. Ordonner les K similarités $\overleftarrow{\mathbf{D}} \in \mathbb{R}^{N_u \times K}$

   $i_1 : \overleftarrow{\mathbf{D}}_{j,i_1} \leq \cdots \leq i_{K-1} : \overleftarrow{\mathbf{D}}_{j,i_{K-1}} \leq i_K : \overleftarrow{\mathbf{D}}_{j,i_K}$

3. Affecter les classes

   $\hat{y}_j = \text{classe\_majoritaire}(y_{i_1}, \cdots, y_{i_K})$ $1 \leq j \leq N_u$

Les K plus proches voisins (K-nn)

Entrées $S_\ell = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N_\ell}$, $S_u = \{\boldsymbol{x}_j\}_{j=1}^{N_u}$ et K

Sortie $\{\hat{y}_j\}_{j=1}^{N_u}$

1. Calculer les similarités $\mathbf{D} \in \mathbb{R}^{N_u \times N_\ell}$

   $\mathbf{D}_{j,i} = d(\boldsymbol{x}_j, \boldsymbol{x}_i)$ $1 \leq j \leq N_u, 1 \leq i \leq N_\ell$

2. Ordonner les K similarités $\overleftarrow{\mathbf{D}} \in \mathbb{R}^{N_u \times K}$

   $i_1 : \overleftarrow{\mathbf{D}}_{j,i_1} \leq \cdots \leq i_{K-1} : \overleftarrow{\mathbf{D}}_{j,i_{K-1}} \leq i_K : \overleftarrow{\mathbf{D}}_{j,i_K}$

3. Affecter les classes

   $\hat{y}_j = \text{classe\_majoritaire}(y_{i_1}, \cdots, y_{i_K})$ $1 \leq j \leq N_u$

What's now ?

Méthodologie

1. Comment évaluer les performances d'un modèle ?
   - Mesures de performance « brutes »
     - ✓ Matrice de confusion
     - ✓ Courbe ROC
   - Procédures d'estimation du risque empirique
     - ⇝ Découpage
     - ⇝ Simulation
     - ⇝ Pénalisation

(Vignette Wikistat)

http://wikistat.fr/pdf/st-m-app-risque.pdf

2. Comment choisir un bon modèle ?
   - ⇝ Inclusion dans les procédures d'estimation du risque empirique

What's now ?

Méthodologie

1. Comment évaluer les performances d'un modèle ?
   - Mesures de performance « brutes »
     - ✓ Matrice de confusion
     - ✓ Courbe ROC
   - Procédures d'estimation du risque empirique
     - ⇝ Découpage
     - ⇝ Simulation
     - ⇝ Pénalisation

(Vignette Wikistat)

http://wikistat.fr/pdf/st-m-app-risque.pdf

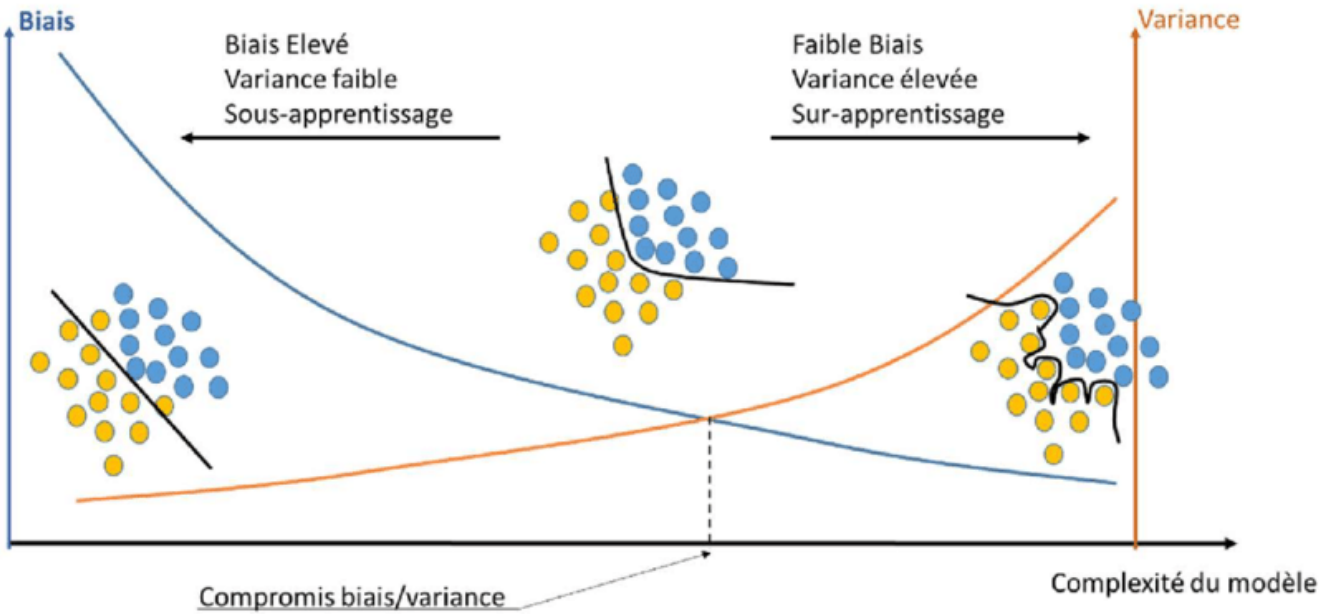
2. Comment choisir un bon modèle ?
   - ⇝ Inclusion dans les procédures d'estimation du risque empirique

Rappel

- Minimisation du RE $h_S^\star = \text{argmin}_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(\boldsymbol{x}_i))$
- Dilemme B vs V $R(h_S^\star) = R(h^- + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h_S^\star) - R(h^\star)]}_{\text{Variance}}$
- Capacité de généralisation de $h_S^\star$ sur S' ? $\rightsquigarrow R(h_{S'}^\star)$

Rappel

- Minimisation du RE $h_S^\star = \text{argmin}_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(\boldsymbol{x}_i))$
- Dilemme B vs V $R(h_S^\star) = R(h^- + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h_S^\star) - R(h^\star)]}_{\text{Variance}}$



Crédits : Pascal Scalar – introduction au DM

- Capacité de généralisation de $h_S^\star$ sur S' ? $\rightsquigarrow R(h_{S'}^\star)$

Rappel

- Minimisation du RE $h_S^\star = \text{argmin}_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, h(\boldsymbol{x}_i))$
- Dilemme B vs V $R(h_S^\star) = R(h^- + \underbrace{[R(h^\star) - R(h^-)]}_{\text{Biais}} + \underbrace{[R(h_S^\star) - R(h^\star)]}_{\text{Variance}}$
- Capacité de généralisation de $h_S^\star$ sur S' ? $\rightsquigarrow R(h_{S'}^\star)$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$

$\{S_A\}_{i=1}^{N_A}$ | $\{S_V\}_{i=1}^{N_V}$ | $\{S_T\}_{i=1}^{N_T}$

$N_A + N_V + N_T = N$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$

$\{S_A\}_{i=1}^{N_A}$	$\{S_V\}_{i=1}^{N_V}$	$\{S_T\}_{i=1}^{N_T}$

$$N_A + N_V + N_T = N$$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$

$\{S_A\}_{i=1}^{N_A}$	$\{S_V\}_{i=1}^{N_V}$	$\{S_T\}_{i=1}^{N_T}$

$$N_A + N_V + N_T = N$$

$h_1 = \text{argmin}_{h \in \mathcal{H}_1} L(S_A)$

$\rightsquigarrow \widehat{R_{A1}}(h_1(S_A))$ $\qquad \rightsquigarrow \widehat{R_{V1}}(h_1(S_V))$

$\vdots \qquad\qquad \vdots$

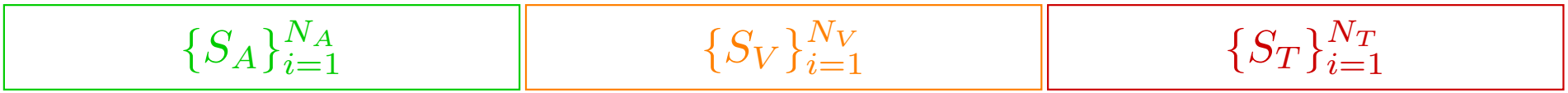
$h_P = \text{argmin}_{h \in \mathcal{H}_P} L(S_A)$

$\rightsquigarrow \widehat{R_{AP}}(h_P(S_A))$ $\qquad \rightsquigarrow \widehat{R_{VP}}(h_P(S_V))$

$p^\star = \text{argmin}_p \widehat{R_{Vp}}(\cdot)$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



$N_A + N_V + N_T = N$

$h_1 = \text{argmin}_{h \in \mathcal{H}_1} L(S_A)$

$\rightsquigarrow \widehat{R_{A1}}(h_1(S_A))$ $\qquad \rightsquigarrow \widehat{R_{V1}}(h_1(S_V))$

$\vdots$ $\qquad \vdots$

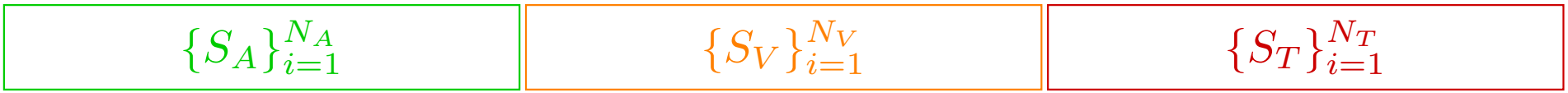
$h_P = \text{argmin}_{h \in \mathcal{H}_P} L(S_A)$

$\rightsquigarrow \widehat{R_{AP}}(h_P(S_A))$ $\qquad \rightsquigarrow \widehat{R_{VP}}(h_P(S_V))$

$p^\star = \text{argmin}_p \widehat{R_{Vp}}(\cdot)$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



$N_A + N_V + N_T = N$

$h_1 = \text{argmin}_{h \in \mathcal{H}_1} L(S_A)$

$\rightsquigarrow \widehat{R_{A1}}(h_1(S_A))$ $\qquad \rightsquigarrow \widehat{R_{V1}}(h_1(S_V))$

$\vdots$ $\qquad \vdots$

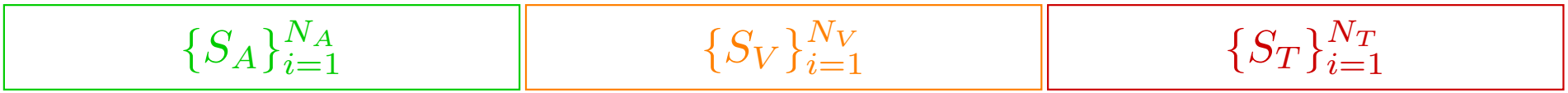
$h_P = \text{argmin}_{h \in \mathcal{H}_P} L(S_A)$

$\rightsquigarrow \widehat{R_{AP}}(h_P(S_A))$ $\qquad \rightsquigarrow \widehat{R_{VP}}(h_P(S_V))$

$p^\star = \text{argmin}_p \widehat{R_{Vp}}(\cdot)$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



$N_A + N_V + N_T = N$

$h_1 = \text{argmin}_{h \in \mathcal{H}_1} L(S_A)$

$\rightsquigarrow \widehat{R_{A1}}(h_1(S_A))$ $\qquad \rightsquigarrow \widehat{R_{V1}}(h_1(S_V))$

$\vdots$ $\qquad \vdots$

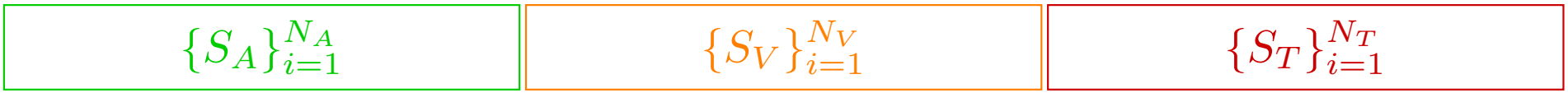
$h_P = \text{argmin}_{h \in \mathcal{H}_P} L(S_A)$

$\rightsquigarrow \widehat{R_{AP}}(h_P(S_A))$ $\qquad \rightsquigarrow \widehat{R_{VP}}(h_P(S_V))$

$p^\star = \text{argmin}_p \widehat{R_{Vp}}(\cdot)$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



$N_A + N_V + N_T = N$

$h_1 = \text{argmin}_{h \in \mathcal{H}_1} L(S_A)$

$\rightsquigarrow \widehat{R_{A1}}(h_1(S_A))$ $\qquad \rightsquigarrow \widehat{R_{V1}}(h_1(S_V))$

$\vdots$ $\qquad \vdots$

$h_P = \text{argmin}_{h \in \mathcal{H}_P} L(S_A)$

$\rightsquigarrow \widehat{R_{AP}}(h_P(S_A))$ $\qquad \rightsquigarrow \widehat{R_{VP}}(h_P(S_V))$

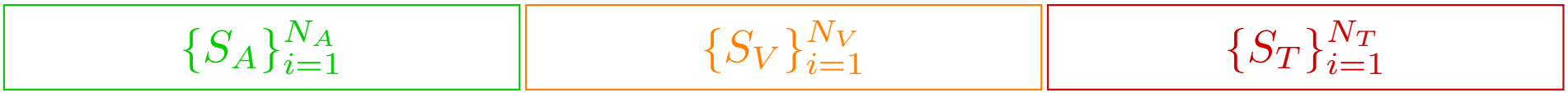
$p^\star = \text{argmin}_p \widehat{R_{Vp}}(\cdot)$

Est. erreur de généralisation

$\rightsquigarrow \widehat{R_{Tp}}(h_{p^\star}(S_T))$

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



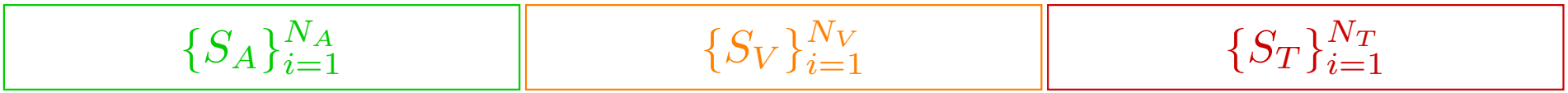
$N_A + N_V + N_T = N$

- Valable pour N suffisamment grand
  - N_A trop petit $\rightsquigarrow$ qualité d'ajustement médiocre
  - $+$
  - N_V ou N_T trop petits $\rightsquigarrow$ variance de l'estimation importante

- Réitérer cette procédure sur B découpages $\rightsquigarrow$ robustesse

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



$N_A + N_V + N_T = N$

- Valable pour N suffisamment grand

  N_A trop petit $\rightsquigarrow$ qualité d'ajustement médiocre

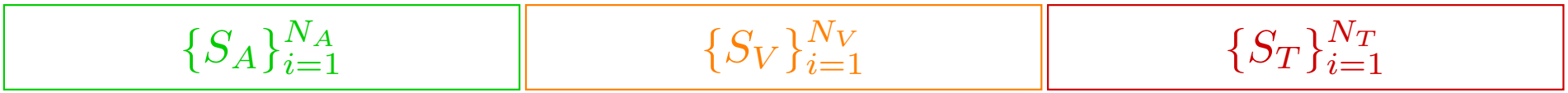
  $+$

  N_V ou N_T trop petits $\rightsquigarrow$ variance de l'estimation importante

- Réitérer cette procédure sur B découpages $\rightsquigarrow$ robustesse
  $\sim$ *Bootstrap* (cf. Vignette Wikistat)

Estimation par découpage aléatoire

Soit $h_p(\boldsymbol{x})$: $\quad h_p = \text{argmin}_{h \in \mathcal{H}_p} L(\{(\boldsymbol{x}_i, y_i)\})$ $\qquad 1 \leq p \leq P$



$$N_A + N_V + N_T = N$$

- Si plusieurs familles de méthodes
  - $\mathcal{H}$: K-nn
  - $\mathcal{G}$: Réseau de neurones
  - $\mathcal{F}$: SVM

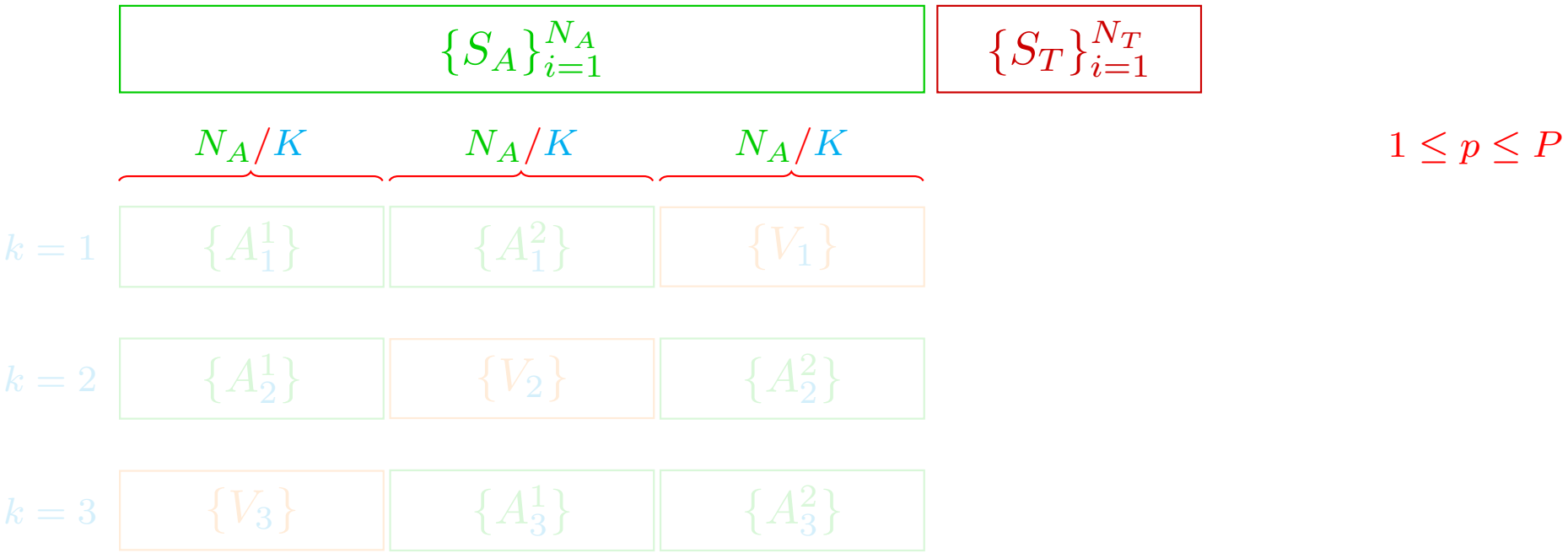
$$\widehat{R_{Tp_h^\star}}(h_{p^\star}(S_T))$$
$$\widehat{R_{Tp_g^\star}}(g_{p^\star}(S_T))$$
$$\widehat{R_{Tp_f^\star}}(f_{p^\star}(S_T))$$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



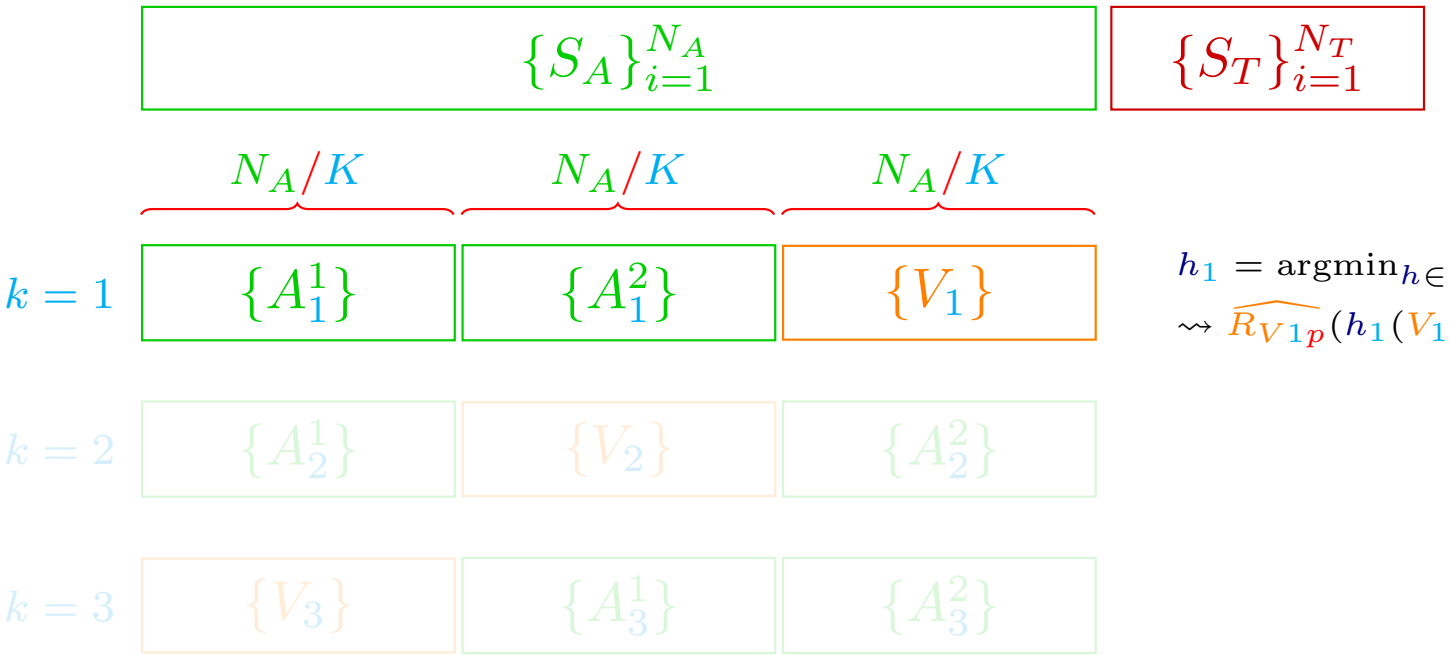
$1 \leq p \leq P$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



$1 \leq p \leq P$

$h_1 = \text{argmin}_{h \in \mathcal{H}_p} L(\{A_1^1\} \cup \{A_1^2\})$

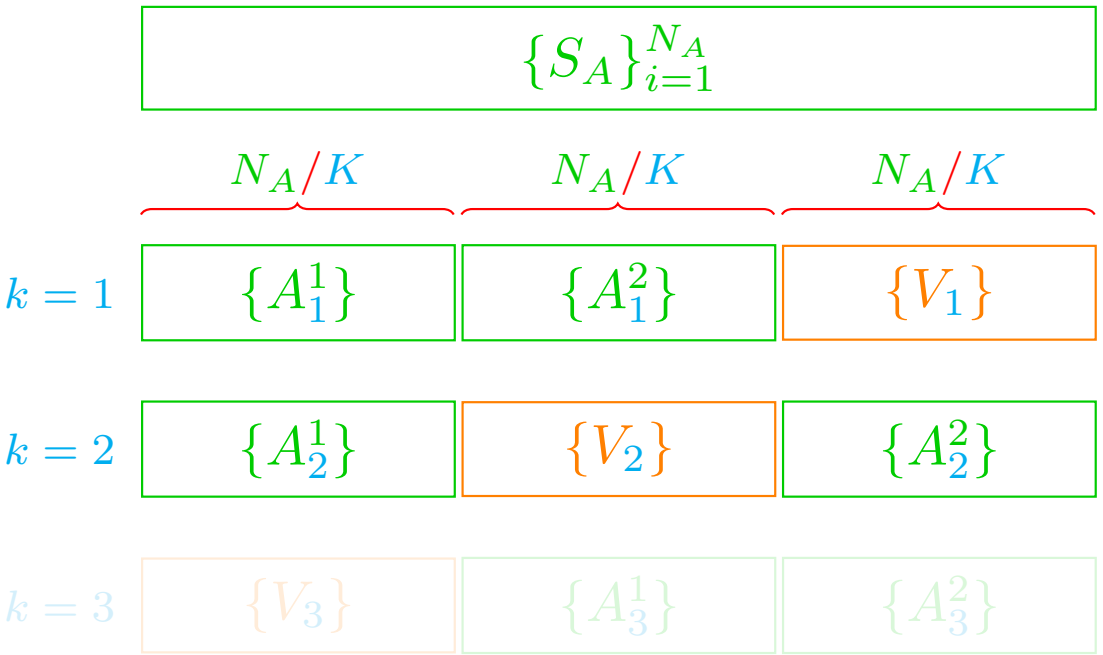
$\rightsquigarrow \widehat{R_{V1p}}(h_1(V_1))$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



$1 \leq p \leq P$

$h_1 = \text{argmin}_{h \in \mathcal{H}_p} L(\{A_1^1\} \cup \{A_1^2\})$

$\rightsquigarrow \widehat{R_{V1p}}(h_1(V_1))$

$h_2 = \text{argmin}_{h \in \mathcal{H}_p} L(\{A_2^1\} \cup \{A_2^2\})$

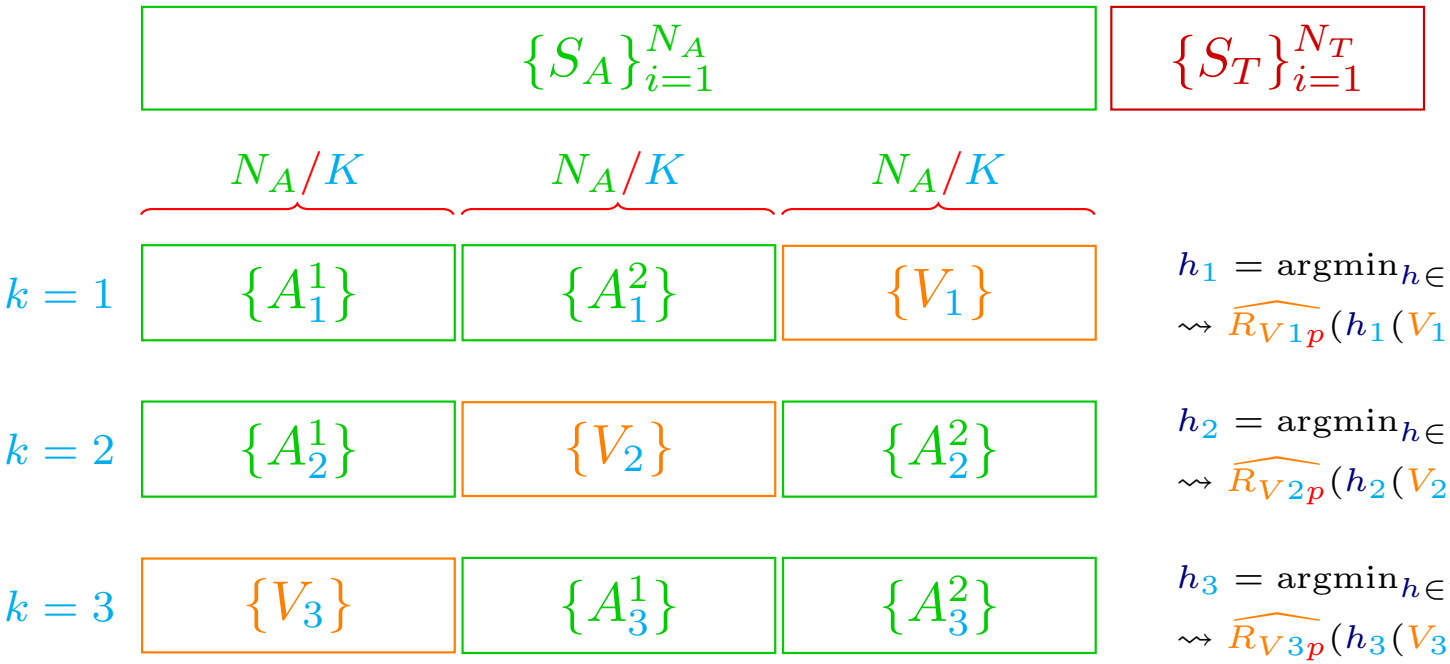
$\rightsquigarrow \widehat{R_{V2p}}(h_2(V_2))$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



$1 \leq p \leq P$

$h_1 = \text{argmin}_{h \in \mathcal{H}_p} L(\{A_1^1\} \cup \{A_1^2\})$
$\rightsquigarrow \widehat{R_{V1p}}(h_1(V_1))$

$h_2 = \text{argmin}_{h \in \mathcal{H}_p} L(\{A_2^1\} \cup \{A_2^2\})$
$\rightsquigarrow \widehat{R_{V2p}}(h_2(V_2))$

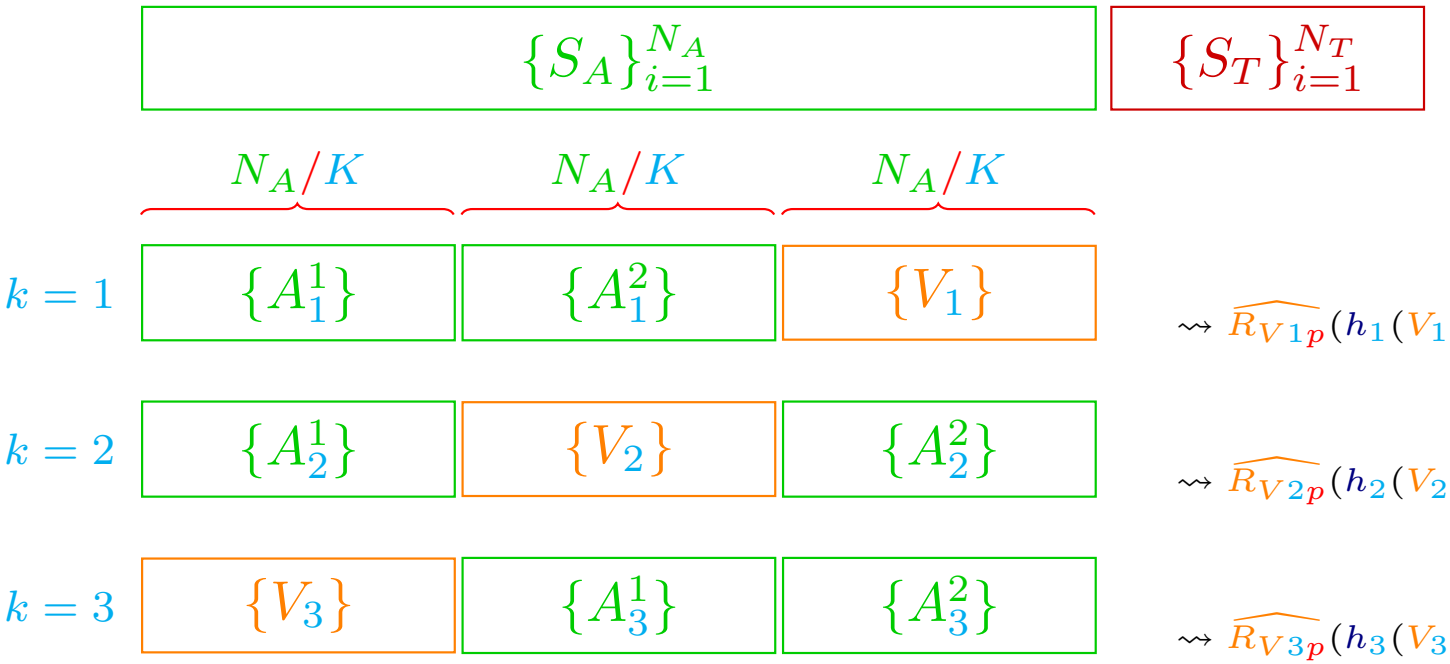
$h_3 = \text{argmin}_{h \in \mathcal{H}_p} L(\{A_3^1\} \cup \{A_3^2\})$
$\rightsquigarrow \widehat{R_{V3p}}(h_3(V_3))$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



$1 \leq p \leq P$

$\rightsquigarrow \widehat{R_{V1p}}(h_1(V_1))$

$\rightsquigarrow \widehat{R_{V2p}}(h_2(V_2))$

$\rightsquigarrow \widehat{R_{V3p}}(h_3(V_3))$

$\rightsquigarrow \widehat{R_{Vp}} = \sum_{k=1}^{K} \widehat{R_{Vkp}}(h_k(V_k))$

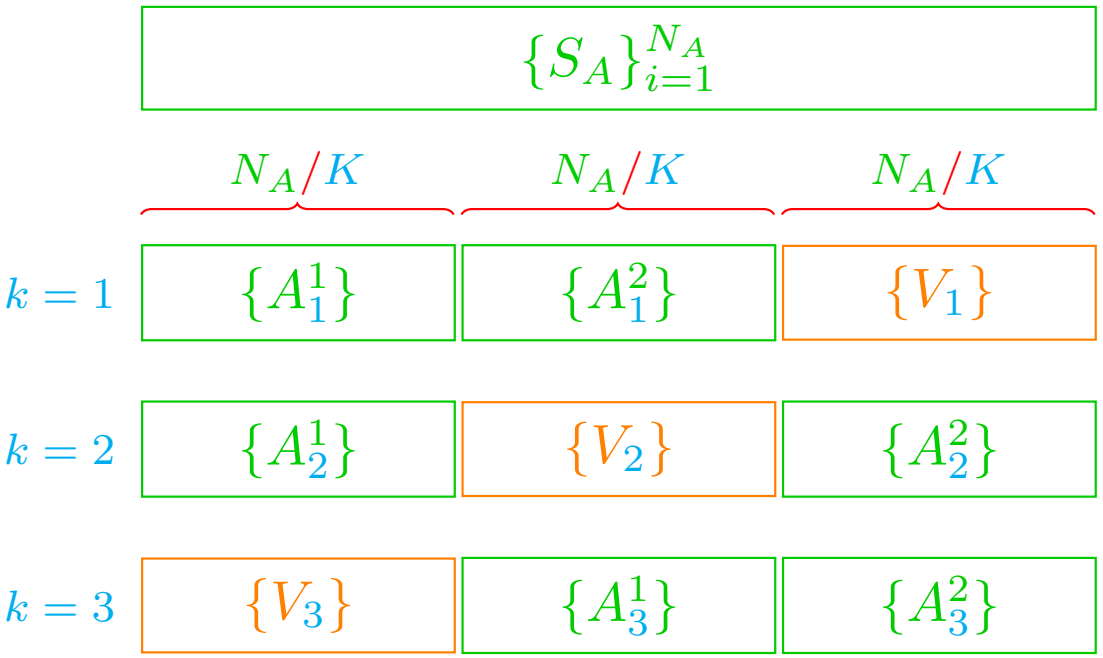
$p^\star = \text{argmin}_p \widehat{R_{Vp}}$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



$1 \leq p \leq P$

$\rightsquigarrow \widehat{R_{V1p}}(h_1(V_1))$

$\rightsquigarrow \widehat{R_{V2p}}(h_2(V_2))$

$\rightsquigarrow \widehat{R_{V3p}}(h_3(V_3))$

$\rightsquigarrow \widehat{R_{Vp}} = \sum_{k=1}^{K} \widehat{R_{Vkp}}(h_k(V_k))$

$p^\star = \text{argmin}_p \widehat{R_{Vp}}$

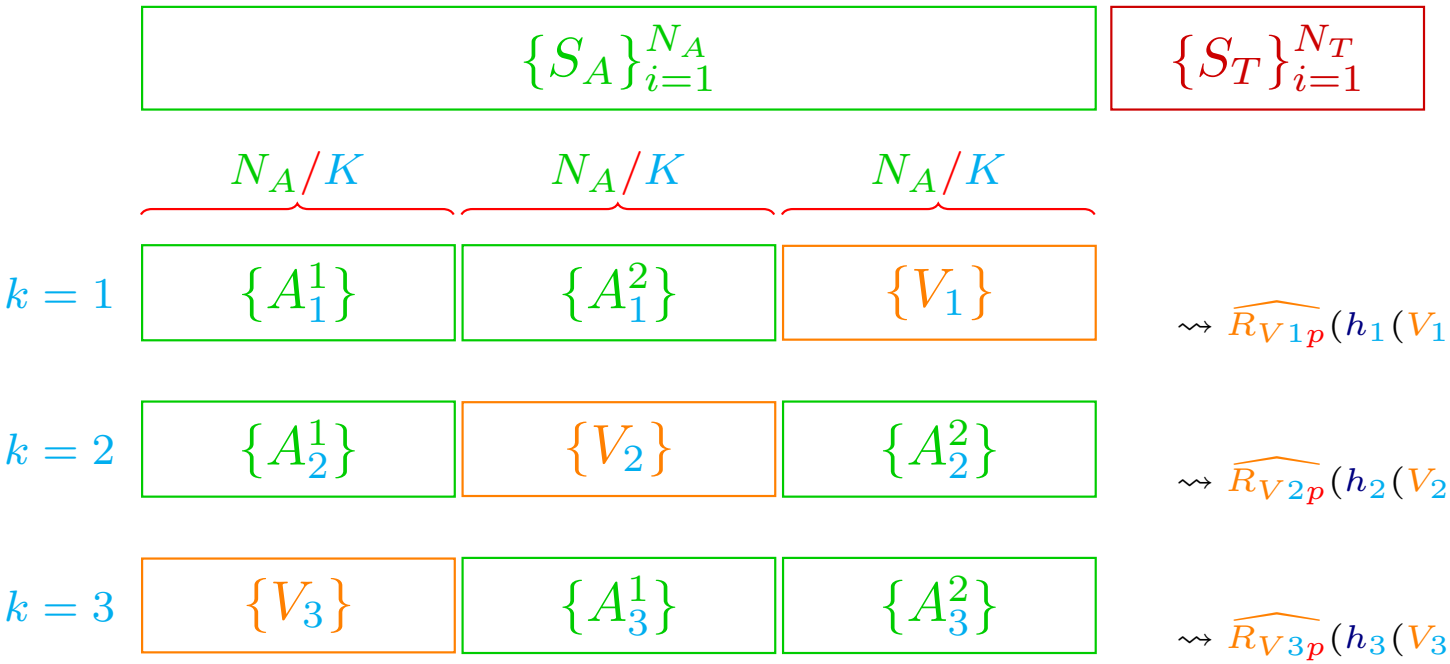
Est. erreur généralisation $h_{p^\star} = \text{argmin}_{h \in \mathcal{H}_{p^\star}} L(S_A)$

Estimation par validation croisée

Principe avec $K = 3$ blocs

K blocs $\neq$ K-nn !

$N_A + N_T = N$



$1 \leq p \leq P$

$\rightsquigarrow \widehat{R_{V1p}}(h_1(V_1))$

$\rightsquigarrow \widehat{R_{V2p}}(h_2(V_2))$

$\rightsquigarrow \widehat{R_{V3p}}(h_3(V_3))$

$\rightsquigarrow \widehat{R_{Vp}} = \sum_{k=1}^{K} \widehat{R_{Vkp}}(h_k(V_k))$

$p^\star = \text{argmin}_p \widehat{R_{Vp}}$

Est. erreur généralisation $h_{p^\star} = \text{argmin}_{h \in \mathcal{H}_{p^\star}} L(S_A)$ $\rightsquigarrow \widehat{R_T}(h_{p^\star}(S_T))$

Estimation par validation croisée

Remarques

- K généralement 5, 10 ou $N_A - 1$ (leave one out)
- Veiller à l'équilibre des classes dans chaque bloc ⇝ stratification
- Réitérer cette procédure sur B découpages ⇝ robustesse ⇝ chronophage

En résumé et en pratique

https://openclassrooms.com/fr/courses/4297211-evaluez-les-performances-dun-modele-de-machine-learning