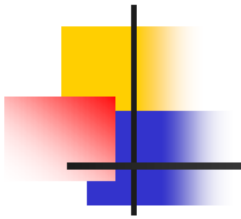
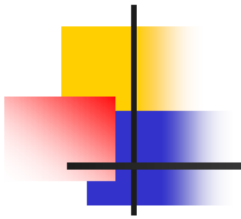


INF7370

Évaluation de l'apprentissage

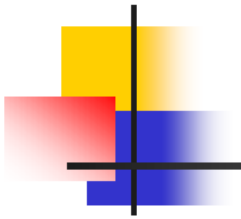
Mohamed Bouguessa



Mise en contexte

L'induction est une forme d'inférence faillible, il faut donc savoir évaluer sa qualité.

- **Questions types:**
  - Quelle est la performance d'un système sur un type de tâche ?
  - Est-ce que mon système est meilleur que l'autre ?
  - Comment dois-je régler mon système ?



Plan

1 – Approches pour évaluer la performance

2 – Métriques pour mesurer la qualité d'un classifieur

3 – La courbe ROC

Approches pour évaluer la performance

L'utilisation d'un échantillon de test

- La méthode la plus simple pour estimer la qualité d'une hypothèse *h* (l'algorithme d'apprentissage) est de diviser l'ensemble des exemples en deux ensembles indépendants: le premier, noté *A*, est utilisé pour l'apprentissage de *h*, le second, noté *T*, sert à mesurer sa qualité. *T* est l'échantillon de test tel que :

$$S = A \bigcup T, et\ A \bigcap T = \phi$$

- La mesure des erreurs commises par *h* sur l'ensemble de test *T* est une estimation la qualité de *h*.

La validation croisée

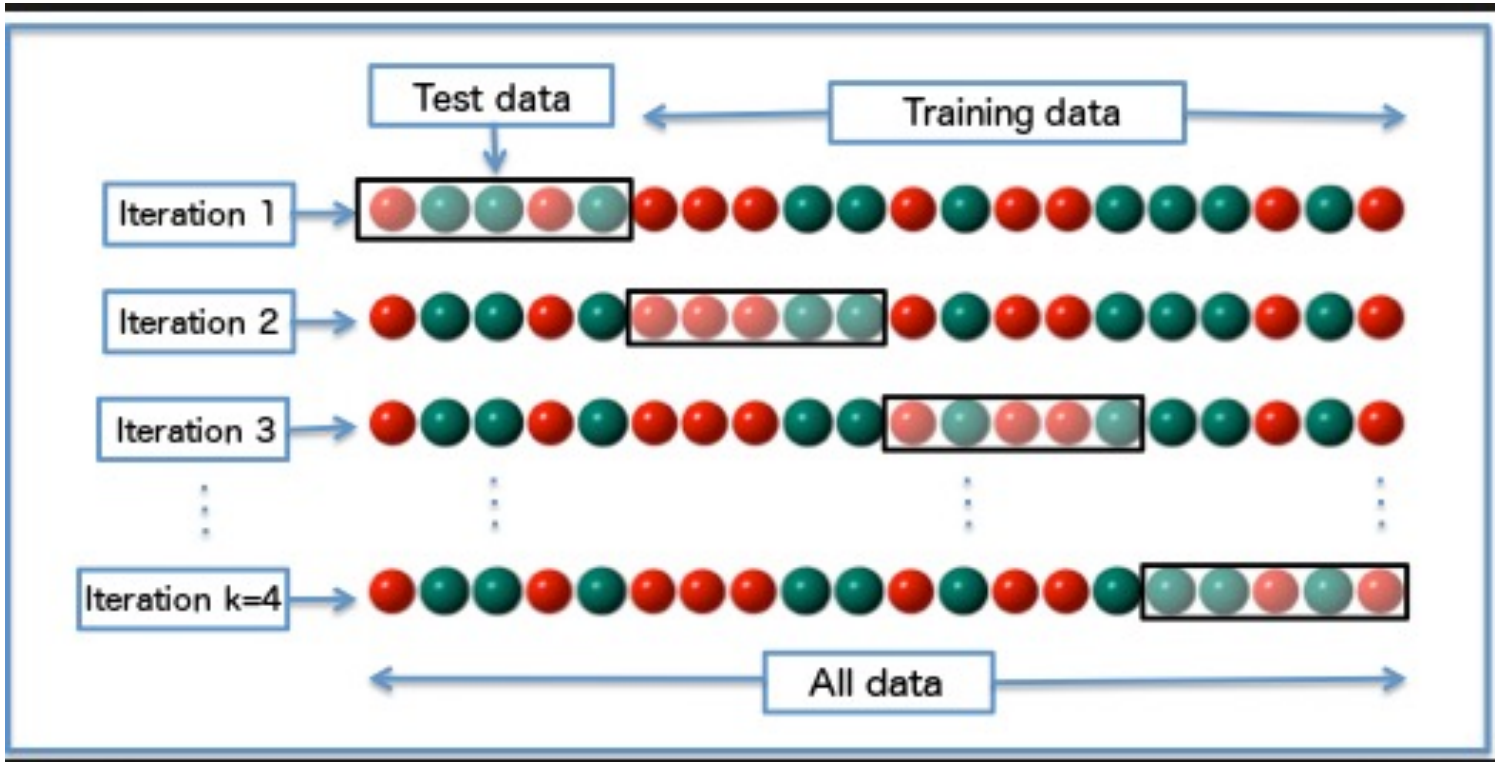
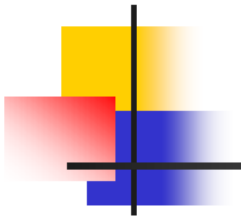


Image du domaine public (Wikipedia)

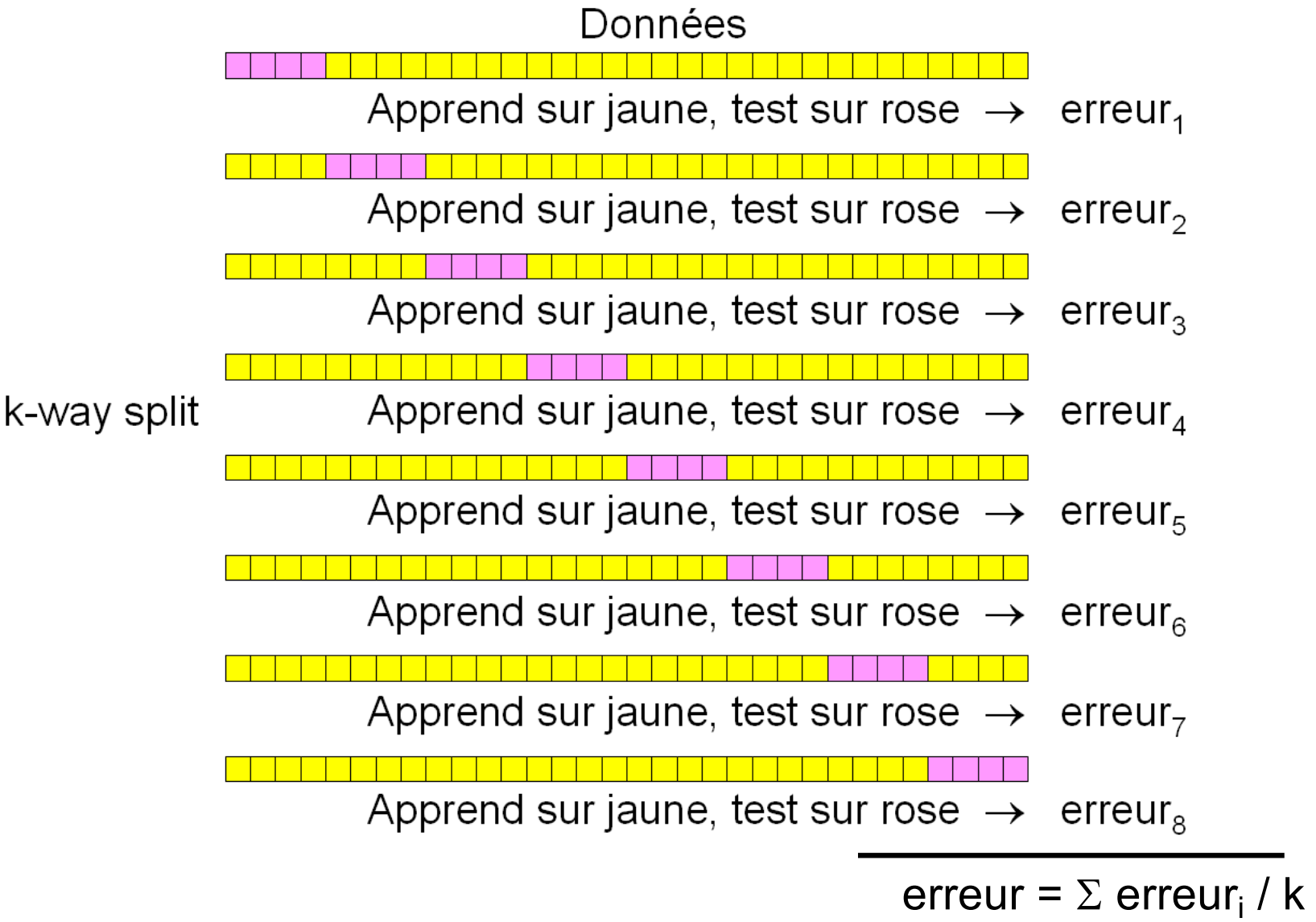


La validation croisée

L'idée de la validation croisée (*k-fold cross-validation*) consiste à :

- Diviser les données d'apprentissage S en k sous-échantillons de tailles égales.
- Retenir l'un de ces échantillons (disons de numéro i). Rouler l'algorithme d'apprentissage sur l'ensemble d'apprentissage.
- Mesurer le taux d'erreur $R_i(h)$ sur l'ensemble de test.
- Recommencer le processus décrit ci-dessus k pour chaque échantillon i.
- L'erreur estimée finale est donnée par la moyenne des erreurs mesurées : $R(h) = \frac{1}{k} \sum_{i=1}^{k} R_i(h)$

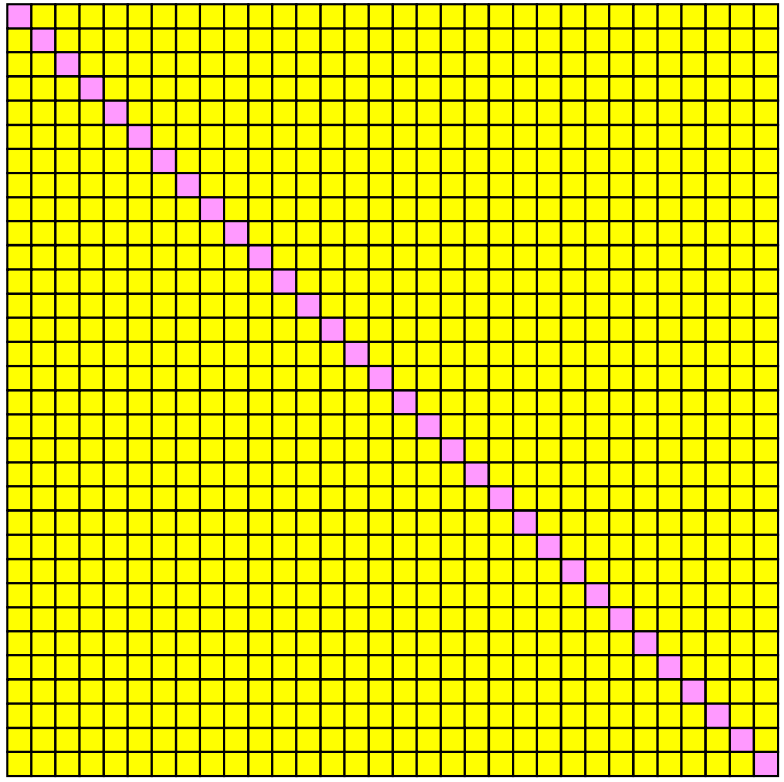
La validation croisée

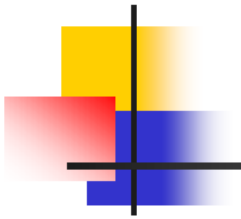


→ La valeur de *k* est généralement comprise entre 5 et 10

Leave-one-out

Lorsque les données disponibles sont très peu nombreuses, il est possible de pousser à l'extrême la méthode de validation croisée en prenant pour k le nombre total d'exemples disponibles ($k=n$). Dans ce cas, on ne retient à chaque fois qu'un seul exemple pour le test, et on répète l'apprentissage k fois pour tous les autres exemples d'apprentissage.





Plan

1 – Approches pour évaluer la performance

2 – Métriques pour mesurer la qualité d'un classifieur

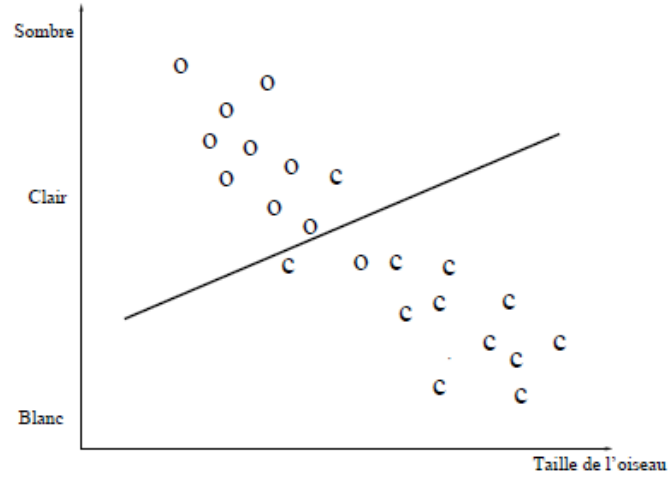
3 – La courbe ROC

Matrice de confusion

- La qualité d'un système de classification est mesurée à l'aide de la matrice de confusion.
- La matrice de confusion *M(i, j)* d'un système de classification est une matrice $C \times C$ dont l'élément générique donne le nombre d'exemples de test de la classe *i* qui ont été classés dans la classe *j*.
- Les colonnes de la matrice de confusion représentent la répartition des objets dans les classes réelles.
- Les lignes représentent la répartition des points dans les classes estimées par un algorithme de classification.

Estimation de la matrice de confusion

❑ Exemple 1 : cas de deux classes



Une règle de décision simple pour séparer les oies des cygnes.

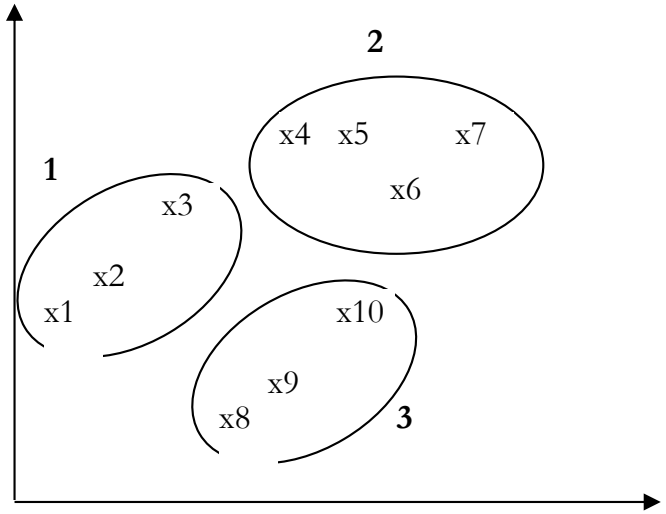
Matrice de confusion d'apprentissage. Erreur empirique : $\frac{1+1}{9+1+1+11} \simeq 9\%$

	O	C
O	9	1
C	1	11

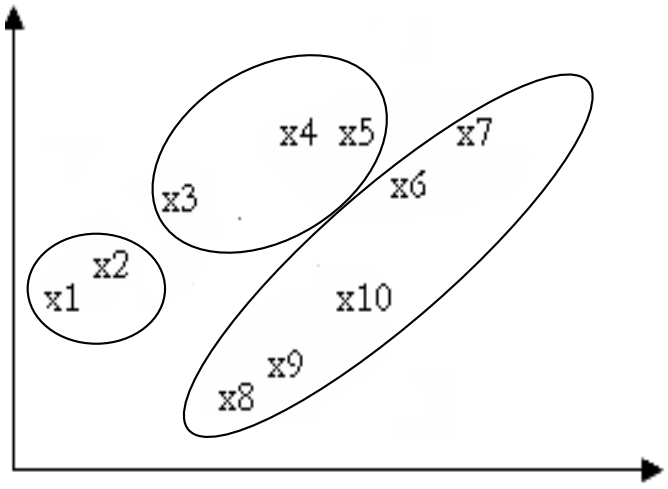
Estimation de la matrice de confusion

❑ Exemple 2 : plusieurs classes

Classes réelles



Classes estimées



Estimation de la matrice de confusion

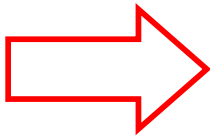
❑ Exemple 2 (suite)

Classes estimées

x1	1
x2	1
x3	2
x4	2
x5	2
x6	3
x7	3
x8	3
x9	3
x10	3

Classes réelles

x1	1
x2	1
x3	1
x4	2
x5	2
x6	2
x7	2
x8	3
x9	3
x10	3



		Classes réelles		
		c1	c2	c3
Classes estimées	C1	2	0	0
	C2	1	2	0
	C3	0	2	3

L'erreur de classification

L'erreur de classification est estimée selon la formule suivante:

$$ER = \frac{\sum_{i}\sum_{j \neq l} n_{ij}}{n}$$

n_{ij} : les différentes valeurs dans la matrice de confusion

l: l'index de la classe réelle c_j avec une valeur maximum n_{ij}

- La valeur de ER est toujours entre 0 et 1.
- Une valeur de ER proche de zéro indique une bonne classification.

Exemple

		Classes réelles		
		c1	c2	c3
Classes estimées	C1	2	0	0
	C2	1	2	0
	C3	0	2	3

$$ER = \frac{(0+0)+(1+0)+(2+0)}{10} = \frac{3}{10} = 0.3$$

Matrice de confusion

- **Un algorithme d'apprentissage qui produit un bon résultat →** Chaque ligne de la matrice de confusion doit avoir une cellule qui contient un plus grand nombre de points comparativement aux autres cellules (de la même ligne). Sinon, si les valeurs de la matrice sont distribuées de façons aléatoires, on peut dire que l'algorithme de classification produit un mauvais résultat.

❑ **Remarque**

- Lorsque la matrice de confusion est très grande (c.-à-d. le nombre de classes est très grand), une lecture simple de cette matrice ne suffit pas pour juger la qualité du résultat.

- Pour cette raison il y a des d'autres mesures, qui sont estimées à partir de la matrice de confusion, pour évaluer la qualité des résultats.

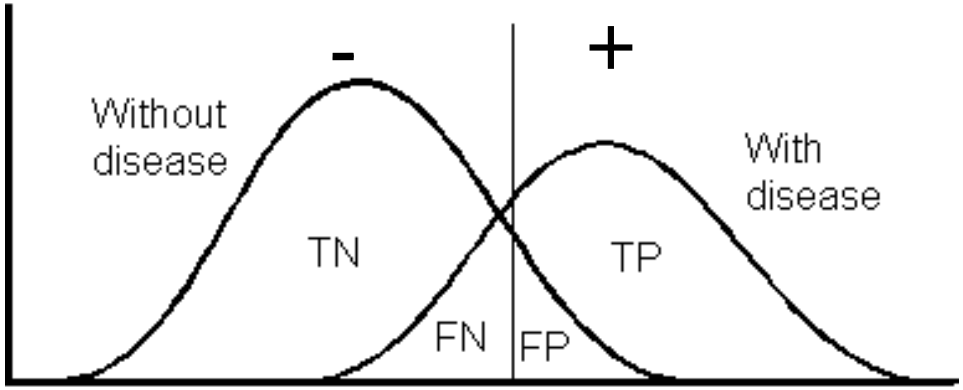
Mesures pour évaluer la performance – Cas de 2 classes

En classification binaire, 4 cas sont possibles

	Classes réelles	
Classes estimées	*TP*	*FP*
	FN	*TN*

- *TP* : nombre de bonnes acceptations ou de vrais positifs (True Positif)
- *FP* : nombre de fausses alarmes ou de faux positifs (Flase Positive)
- *FN* : nombre de faux rejets ou de faux négatifs (False Negative)
- *TN* : nombre de rejets corrects ou de vrais négatifs (True Negative)

Exemple



Réel / Estimé	+	-
+	***TP***	***FP***
-	***FN***	***TN***

Mesures pour évaluer la performance

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

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

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5	1
	Negative	2	2

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

- La précision permet de répondre à la question suivante : Quelle proportion d'identifications positives était effectivement correcte ?
- Le rappel permet de répondre à la question suivante : Quelle proportion de résultats positifs réels a été identifiée correctement ?
- Remarque : Le Rappel (en anglais, recall) est également appelé ***sensibilité*** ou encore TPR (True Positive Rate).

Mesures pour évaluer la performance

$$F_1 = \frac{2 \times Précision \times Rappel}{Précision + Rappel}$$

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

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

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5	1
	Negative	2	2

		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

- La valeur du score F_1 est toujours entre 0 et 1.
- Une valeur proche de 1 indique la classification est de bonne qualité.

Mesures pour évaluer la performance

$$F_1 = \frac{2 \times Précision \times Rappel}{Précision + Rappel}$$

$$\text{Précision} = \frac{TP}{TP+FP} \qquad \text{Rappel} = \frac{TP}{TP+FN}$$

Le score F_1 est la moyenne harmonique de la précision et du rappel. Alors que la moyenne ordinaire traite de la même manière toutes les valeurs, la moyenne harmonique donne plus de poids aux faibles valeurs. Par conséquent, le classificateur n'obtiendra un bon score F_1 que si son Rappel et sa Précision sont élevés.*

*Source: Aurélien Géron. Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets. Dunod, 2019.

Exemple*

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

Classe Cat:

$$\text{Précision} = \frac{TP}{TP+FP} = \frac{4}{4+6+3} = 0.3076$$

$$\text{Rappel} = \frac{TP}{TP+FN} = \frac{4}{4+1+1} = 0.6666$$

*source: https://towardsdatascience.com/

Exemple*

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6

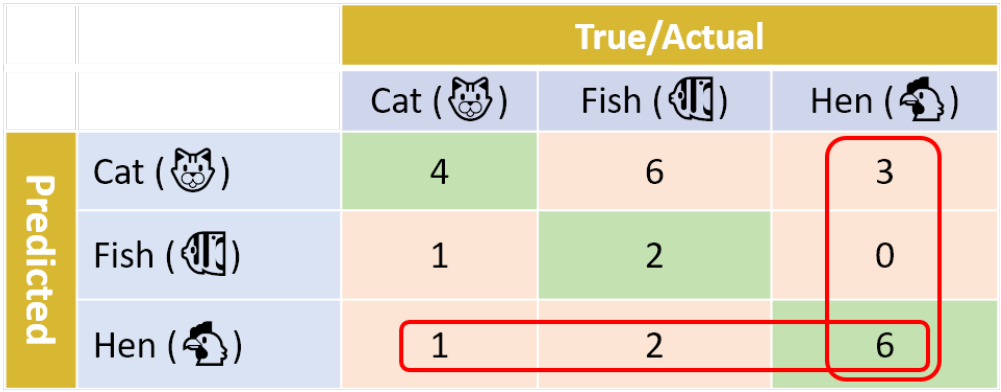
Classe Fish: $\text{Précision} = \frac{TP}{TP+FP} = \frac{2}{2+1+0} = 0.6666$

$\text{Rappel} = \frac{TP}{TP+FN} = \frac{2}{2+6+2} = 0.2$

*source: https://towardsdatascience.com/

Exemple*

		True/Actual		
		Cat (🐱)	Fish (🐟)	Hen (🐔)
Predicted	Cat (🐱)	4	6	3
	Fish (🐟)	1	2	0
	Hen (🐔)	1	2	6



Classe Hen: $\text{Précision} = \frac{TP}{TP+FP} = \frac{6}{6+3+0} = 0.6666$

$\text{Rappel} = \frac{TP}{TP+FN} = \frac{6}{6+2+1} = 0.6666$

*source: https://towardsdatascience.com/

Remarque

- Le score F_1 favorise les classificateurs ayant une précision et un rappel similaires. Ce n'est pas toujours ce que nous voulions :
- Dans certains contextes, nous nous soucions essentiellement de la Précision. Exemple : un classifieur de détection de vidéos sans danger pour les enfants. On a, probablement, tendance a préféré un classifieur qui rejette beaucoup de bonnes vidéos (faible rappel), mais conserve uniquement des vidéos sans danger (haute précision).

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

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

Réel / Estimé	+	-
+	TP	FP
-	FN	TN

+ : vidéos sans dangers pour les enfants

- : autres

*Source: Aurélien Géron. Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets. Dunod, 2019.

Remarque

- Le score F_1 favorise les classificateurs ayant une précision et un rappel similaires. Ce n'est pas toujours ce que nous voulions :
- D'autres contextes c'est surtout le Rappel qui nous intéresse. Exemple : détection des voleurs à l'étalage sur des images surveillance. Probablement, on peut se contenter d'un classifieur avec une précision de 30%, mais un rappel de 99% (bien évidemment, les agents de sécurité recevront quelques fausses alertes, mais presque tous les vols à l'étalage seront interceptés).*

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

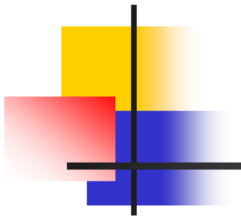
$$\text{Rappel} = \frac{TP}{TP+FN}$$

Réel / Estimé	+	-
+	TP	FP
-	FN	TN

+ : vol à l'étalage

- : autres

*Source: Aurélien Géron. Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets. Dunod, 2019.



Plan

1 – Approches pour évaluer la performance

2 – Métriques pour mesurer la qualité d'un classifieur

3 – La courbe ROC

Faux positifs et faux négatifs

Réel / Estimé	+	-
+	TP	FP
-	FN	TN

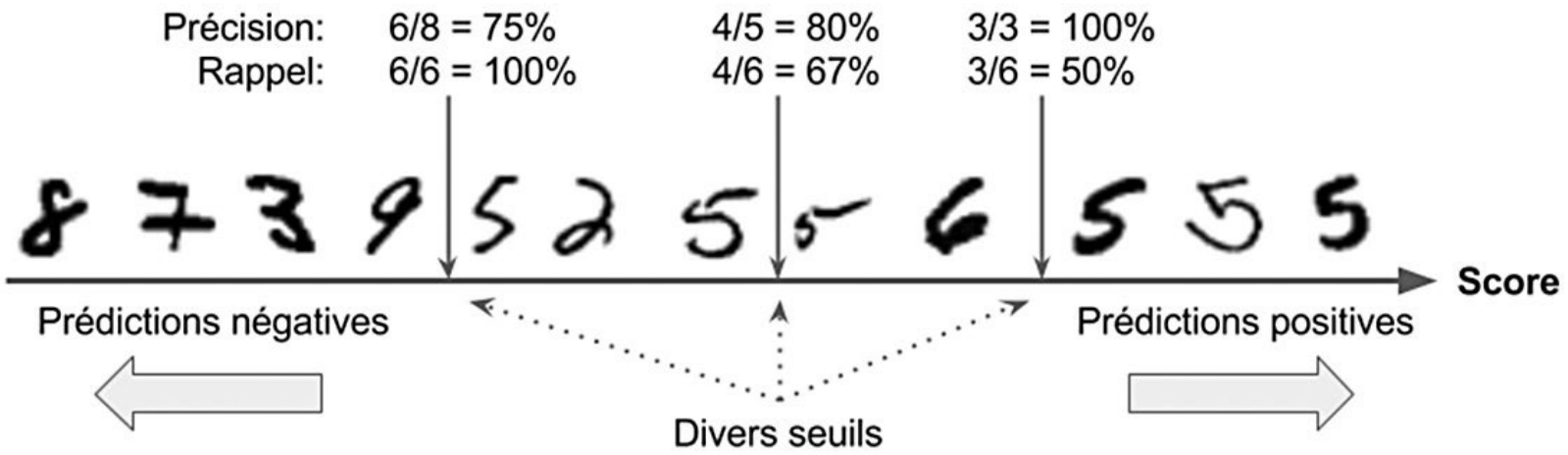
		True/Actual	
		Positive (🐶)	Negative
Predicted	Positive (🐶)	5 (TP)	1 (FP)
	Negative	2 (FN)	2 (TN)

- Erreur de type 1: **faux positifs**
  - Probabilité d'accepter l'hypothèse alors qu'elle est fausse
- Erreur de type 2 : **faux négatifs**
  - Probabilité de rejeter l'hypothèse alors qu'elle est vraie

Dans un contexte de prise de décision, il est parfois utile dans l'évaluation des performances de prendre en compte non seulement un taux d'erreur, mais aussi les taux de « faux positifs » et de « faux négatifs ». On peut préférer avoir un taux d'erreur moins bon si cela permet de réduire le type d'erreur le plus couteux.

Exemple*

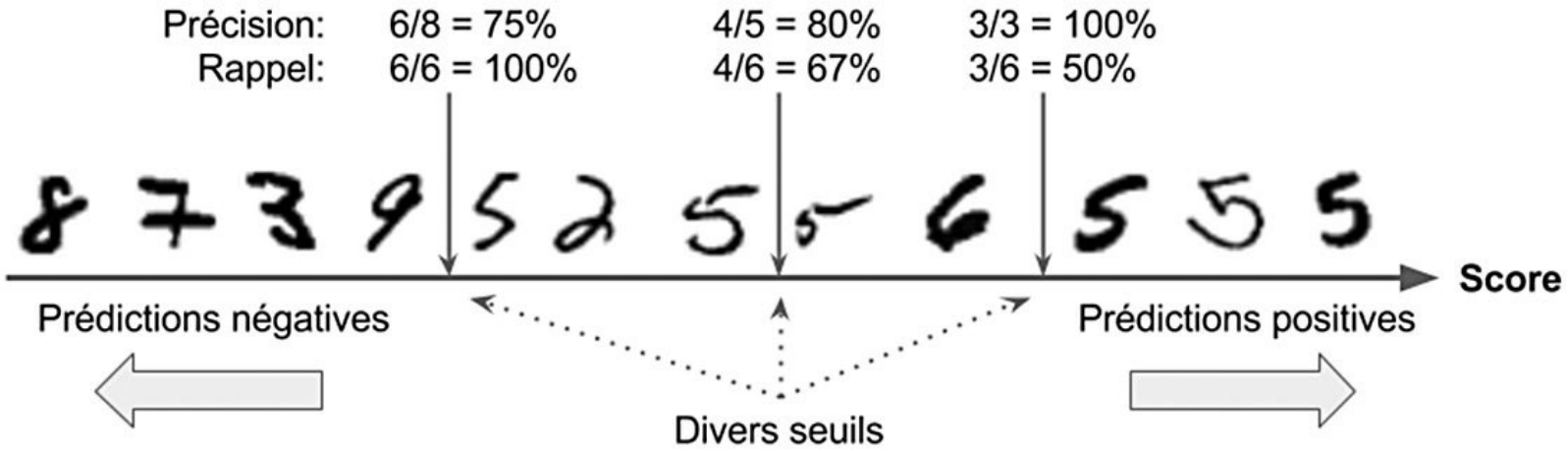
- « détecteur de 5 » un *classifieur binaire* capable d'effectuer la distinction entre deux classes uniquement, 5 et non-5.
- Pour chaque observation, le classifieur calcule un score basé sur une *fonction de décision*. Si ce score est supérieur à un certain seuil, il affecte l'observation à la classe positive, sinon il l'affecte à la classe négative.



*Source: Aurélien Géron. Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets. Dunod, 2019.

Exemple* (suite)

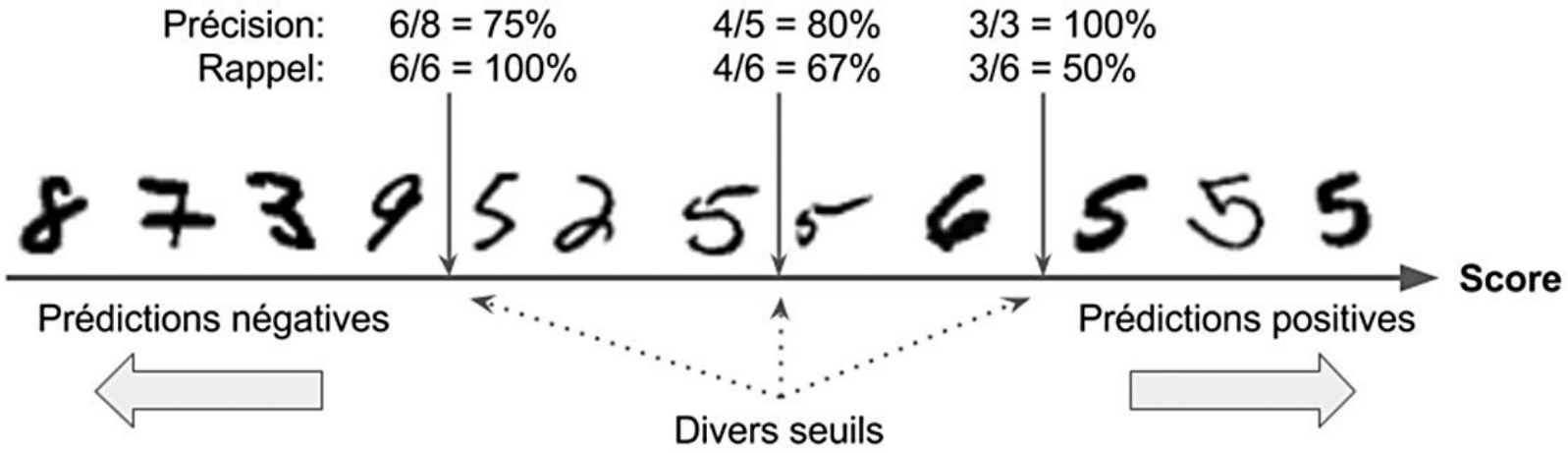
Cette figure présente un certain nombre de chiffres s'échelonnant depuis le score le plus faible à gauche jusqu'au score le plus élevé à droite. Supposons que le *seuil de décision* soit positionné au niveau de la flèche centrale (entre les deux 5) : vous trouverez 4 vrais positifs (vrais 5) à droite de ce seuil, et un faux positif (en fait un 6). Par conséquent, avec ce seuil, la précision et de 80 % (4 sur 5). Mais sur 6 véritables 5, le classificateur n'en détecte que 4, soit un rappel de 67 % (4 sur 6).*



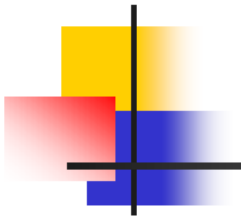
*Source: Aurélien Géron. Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets. Dunod, 2019.

Exemple* (suite et fin)

Si on augmente le seuil (en déplaçant la flèche vers la droite), le faux positif (le 6) devient maintenant un vrai négatif, accroissant de ce fait la précision (jusqu'à 100 % dans ce cas), mais un vrai positif devient maintenant un faux négatif, faisant baisser ainsi le rappel à 50 %. Inversement, abaisser le seuil accroît le rappel et réduit la précision.*



*Source: Aurélien Géron. Machine Learning avec Scikit-Learn : Mise en œuvre et cas concrets. Dunod, 2019.



Courbe ROC

Comment arbitrer entre ces types d'erreurs : les « faux positifs » et les « faux négatifs » ?

- La courbe d'efficacité du récepteur (Receiver Operating Characteristic ou ROC) permet de régler ce compromis.
- La courbe ROC est avant tout définie pour les problèmes à deux classes (les positifs et les négatifs).
- Elle met en relation dans un graphique les taux de faux positifs (en abscisse) et les taux de vrais positifs (en ordonnée).

Courbe ROC

- La courbe ROC croise le taux de faux positifs (False Positive Rate ou FPR) et le taux de vrais positifs (True Positive Rate ou TPR – un autre nom pour le Rappel)
- Chaque classifieur produit un point (taux FP, taux TP) dans la courbe.
- Taux de FP (False Positive Rate)

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

- Taux de TP (True Positive Rate)

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

Réel / Estimé	+	-
+	TP	FP
-	FN	TN

Courbe ROC

- Chaque classifieur produit un point (taux FP, taux TP) dans la courbe.

Réel / Estimé	+	-
+	TP	FP
-	FN	TN

- Taux de FP (False Positive Rate)

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

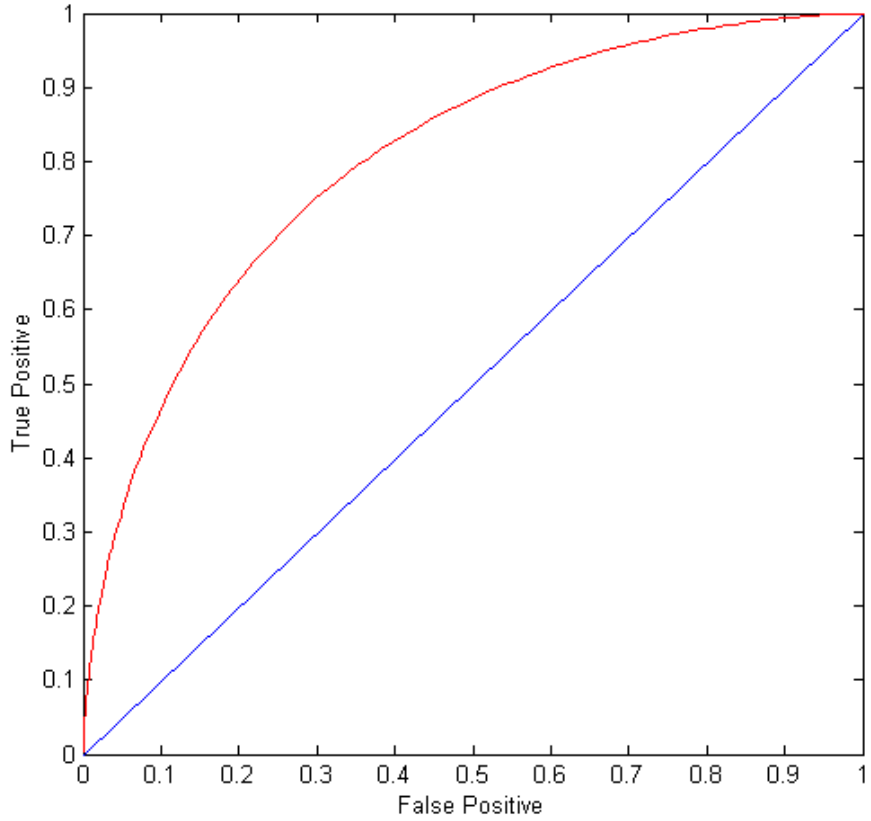
Remarque : Le taux de faux positifs (*False Positive Rate* ou *FPR*) est le pourcentage d'observations négatives qui sont incorrectement classées comme positives. Il est égal à 1 moins le *taux de vrais négatifs*, qui est le pourcentage d'observations négatives qui sont correctement classées comme négatives. Le taux de faux négatifs (*True Negative Rate* ou TNR) est aussi appelé *spécificité*. Par conséquent la courbe ROC croise sensibilité et 1 – spécificité.

Courbe ROC

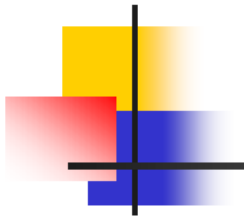
Quelques points importants dans la courbe

(FPR, TPR):

- (0, 0) prédit toujours négatif
- (1, 1) prédit toujours positif
- (0, 1) classification idéale
- Ligne diagonale (ligne de hasard): classification aléatoire (pertinence = 0.5)
- Le triangle sous la diagonale est pire que le hasard



→ Comment construire la courbe ROC?



Cadre

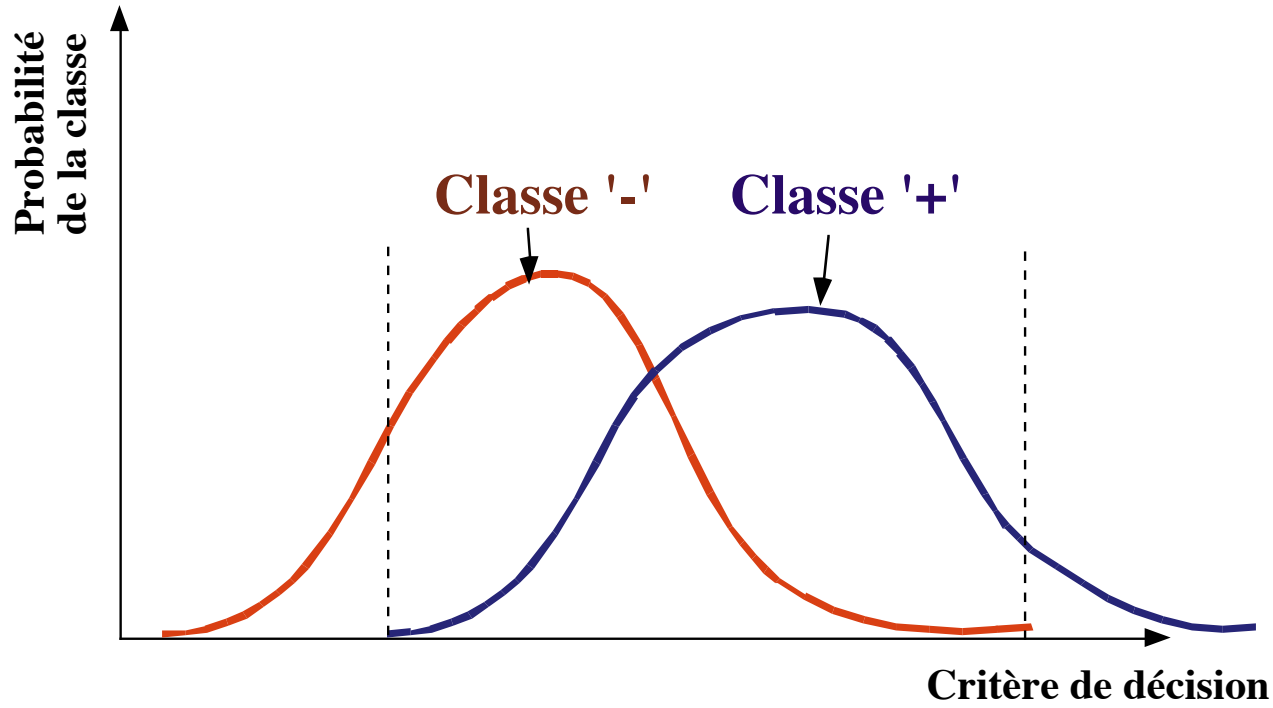
- Les classifieurs discrets (arbre de décision par exemple) renvoie seulement une classe de décision, et donc une seule matrice de confusion :→ donc un seul point de l'espace ROC.
- D'autres classifieurs ne renvoient pas seulement une classe de décision, mais un score ou une probabilité qui représente le degré d'appartenance d'un exemple à une classe spécifique. Nous considérons ce type de classificateurs pour la construction de la courbe ROC.

→ Un classificateur à score peut être utilisé avec un **seuil** donne un classificateur discret : Si le score est supérieur à un seuil, on prédit la classe +, sinon la classe -.

Construction de la courbe ROC

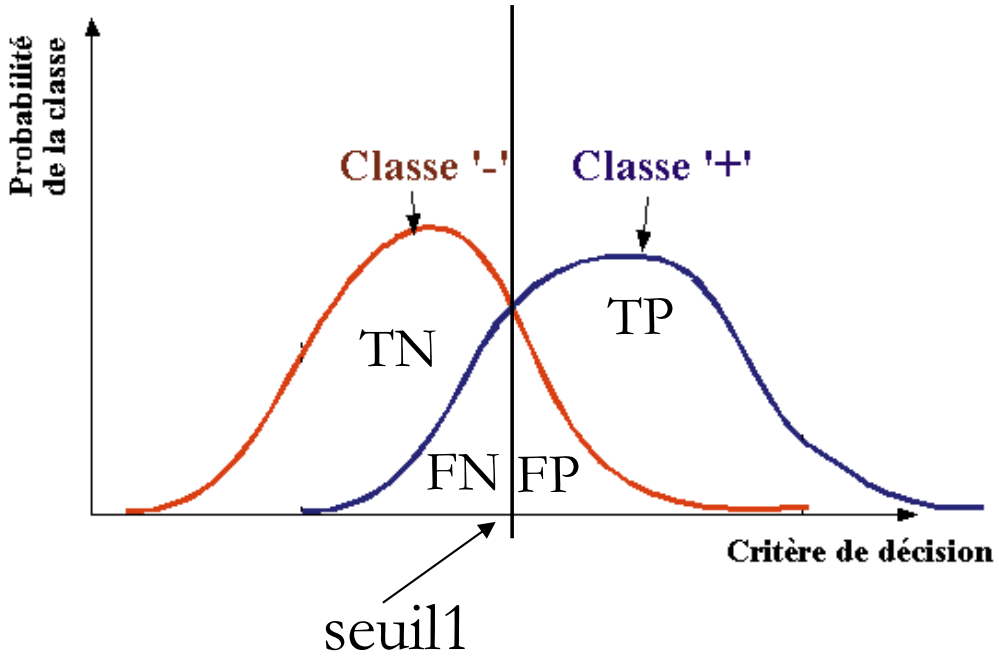
- La courbe ROC est définie pour les problèmes de deux classes (ex. classification des patients en deux classes: saint et malade).

→ On peut donc établir un graphique qui représente la probabilité d'appartenir à une classe spécifique.

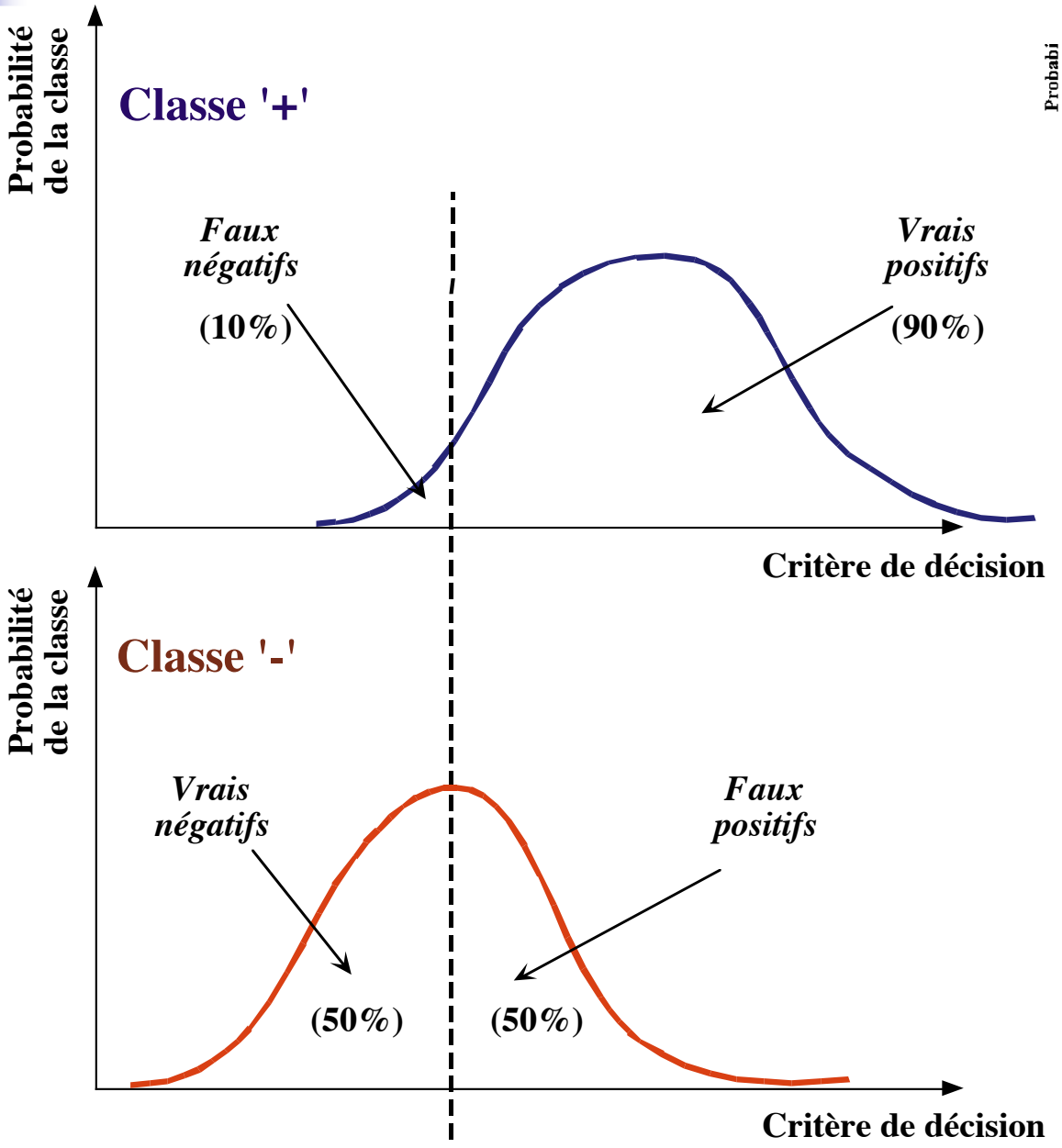


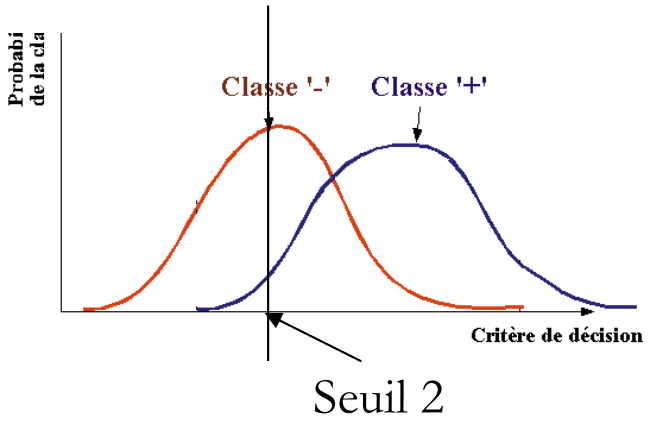
Construction de la courbe ROC

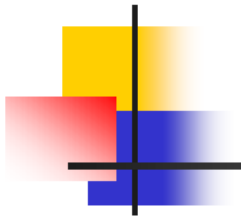
L'étape suivante consiste à chercher un seuil pour séparer les deux classes. Pour ce faire, on fait varier la valeur du seuil. Pour chacune de ces variations on calcule la le taux de FP et le taux de TP. Chaque valeur du seuil est associée donc à un couple (taux FP, taux TP) qui représente à son tour un point de la courbe ROC.



Construction de la courbe ROC



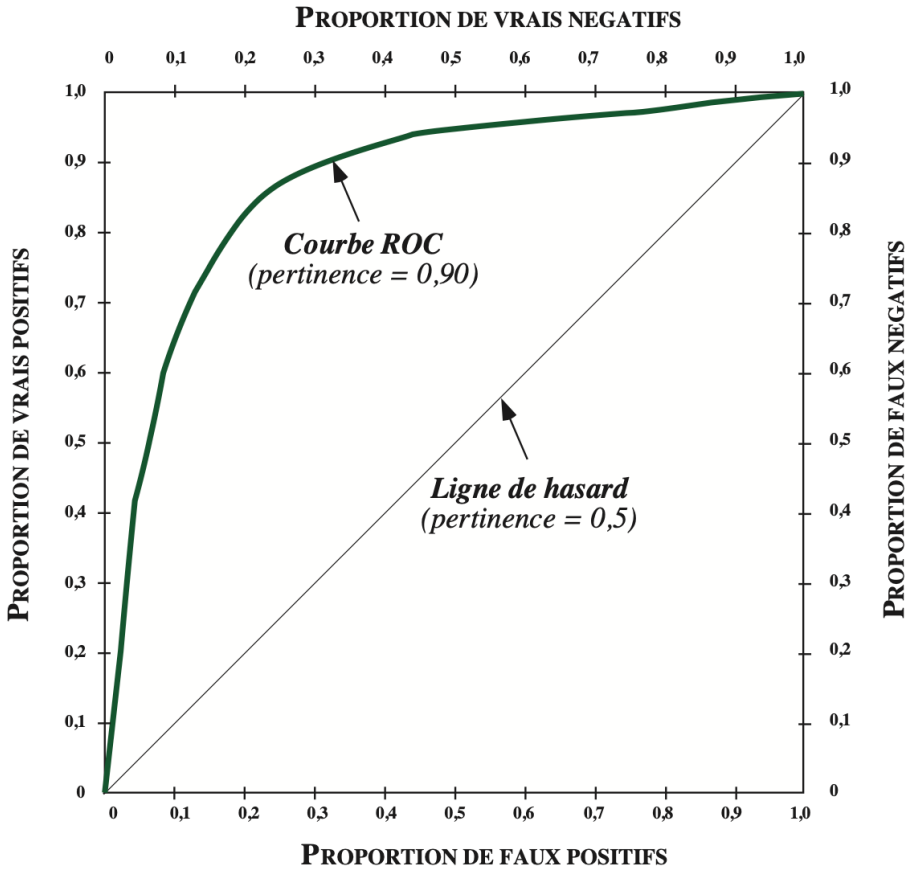


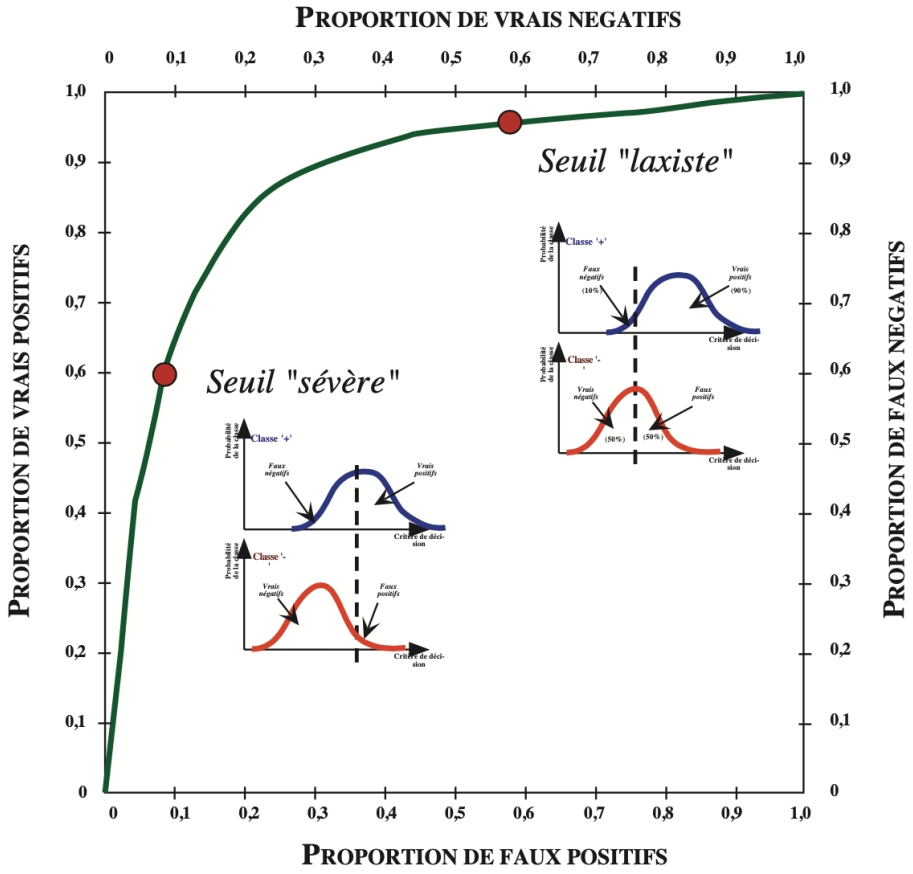


Construction de la courbe ROC

- La dernière étape correspond à la construction de la courbe. Pour chaque seuil, on rapporte la proportion des TP en fonction de celles des FP.
- Si l'on obtient une droite, on doit conclure que le test a 50% de construire un bon diagnostique.
- Plus la courbe s'incurve vers le haut plus le test est pertinent. On peut donc utiliser la courbe pour décider quel est meilleur seuil. Il s'agira du seuil où la courbe ROC montre un point d'inflexion.

Construction de la courbe ROC

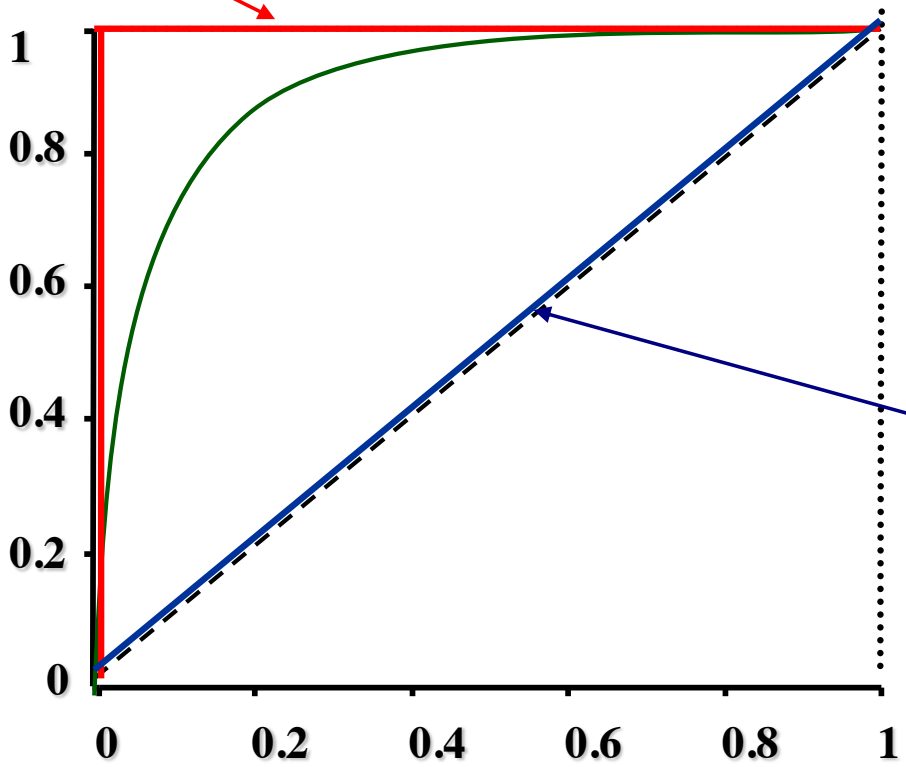




*Source: A. Cornuéjolus et al. Apprentissage artificiel : deep learning, concepts et algorithmes. Eyrolles, 2018.

Interprétation de la courbe ROC

Séparation parfaite

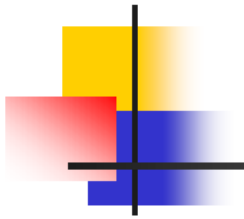


Aucune séparation

Exemple

Supposons que nous avons un ensemble d'apprentissage avec 10 exemples répartis en deux classes {+, -}. Soit h(x)→{+, -} un classifieur qui retourne un score/une probabilité. Le tableau suivant les scores ainsi que la classe réelle de chaque exemple.

Objet	Score	Classe réelle
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	+
5	0.83	-
6	0.80	-
7	0.76	-
8	0.53	+
9	0.43	-
10	0.25	+



Exemple (suite)

Pour construire la courbe ROC, on doit :

1. Trier les scores;
2. Établir des seuils de séparation entre les différentes valeurs des scores → Appliquer un seuil à chaque valeur du score (apply a threshold at each unique value of the score)
3. Calculer le taux de FP et le taux de TP pour chaque intervalle (chaque intervalle est défini par un seuil).

Exemple (suite)

Class	+	-	+	-	-	-	+	-	+	+	
seuil >=	0.25	0.43	0.53	0.76	0.80	0.83	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

← Score

Courbe ROC: