Chapitre : Evaluation de l'apprentissage

Plan

1. Introduction
2. Matrice de confusion
3. Courbe ROC
4. Erreurs de régression
5. Généralisation et sur-apprentissage

Introduction

- 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 ?

Introduction

Types de mesures de performance

Il existe de nombreuses façons d'évaluer la performance prédictive d'un modèle d'apprentissage supervisé. les principaux critères utilisés

SVM

```
Correctly Classified Instances         117               70.9091 %
Incorrectly Classified Instances        48               29.0909 %
Kappa statistic                          0.3071
Mean absolute error                      0.2909
Root mean squared error                  0.5394
Relative absolute error                 62.6804 %
Root relative squared error            112.1168 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.617      0.718     0.895     0.797      0.639    good
                 0.383     0.105      0.676     0.383     0.489      0.639    bad
Weighted Avg.    0.709     0.431      0.703     0.709     0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

Naive Bayes

```
Correctly Classified Instances         103               62.4242 %
Incorrectly Classified Instances        62               37.5758 %
Kappa statistic                          0.1995
Mean absolute error                      0.3793
Root mean squared error                  0.5316
Relative absolute error                 81.7353 %
Root relative squared error            110.5048 %
Total Number of Instances              165

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.686     0.483      0.713     0.686     0.699      0.674    good
                 0.517     0.314      0.484     0.517     0.5        0.674    bad
Weighted Avg.    0.624     0.422      0.63      0.624     0.627      0.674

=== Confusion Matrix ===

  a  b   <-- classified as
 72 33 |  a = good
 29 31 |  b = bad
```

Introduction

Types de mesures de performance

- Test set: 105 good, 60 bad
  - NB Accuracy 62.4%
  - SVM Accuracy 70.1%

SVM

Classified as

good	bad	
94	**11**	good
37	**23**	bad

Act. Class

Naive Bayes

Classified as

good	bad	
72	**33**	good
29	**31**	bad

Act. Class

Introduction

Types de mesures de performance

- Test set: 105 good, 60 bad
  - NB Accuracy 62.4%
  - SVM Accuracy 70.1%

SVM

Classified as

good	bad	
94	**11**	good
37	**23**	bad

Act. Class

SVM biased toward majority class

Naive Bayes

Classified as

good	bad	
72	**33**	good
29	**31**	bad

Act. Class

What if this is important?

Introduction

Types de mesures de performance

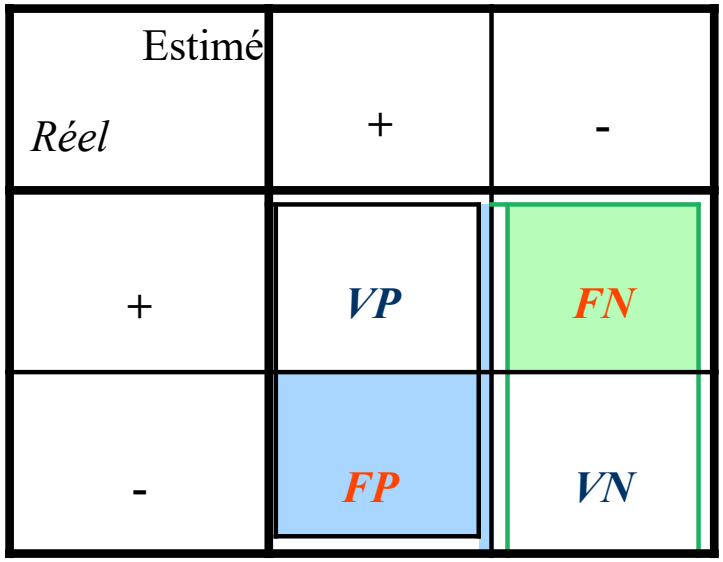
```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.895     0.617      0.718     0.895      0.797      0.639    good
                 0.383     0.105      0.676     0.383      0.489      0.639    bad
Weighted Avg.    0.709     0.431      0.703     0.709      0.685      0.639

=== Confusion Matrix ===

  a  b   <-- classified as
 94 11 |  a = good
 37 23 |  b = bad
```

Matrice de confusion

- Étant donné un problème de classification, on appelle matrice de confusion une matrice **M** contenant autant de lignes que de colonnes que de classes, et dont l'entrée M_{ck} est le nombre d'exemples de la classe **c** pour laquelle l'étiquette **k** a été prédite

Estimé / *réel*	+	-
+	*VP*	*FN*
-	*FP*	*VN*

Matrice de confusion



- VP: vrais positifs (true positives) les exemples positifs correctement classifiés
- FP: faux positifs (false positives) les exemples négatifs étiquetés positifs par le modèle ;
- et réciproquement pour les vrais négatifs VN (true negatives) et les faux négatifs FN (false negatives).
- On note généralement par TP le nombre de vrais positifs, FP le nombre de faux positifs, TN le nombre de vrais négatifs et FN le nombre de faux négatifs.

- Les faux positifs sont aussi appelés fausses alarmes.

Matrice de confusion

Il est possible de dériver de nombreux critères d'évaluation à partir de la matrice de confusion.

- On appelle **rappel** (recall), ou **sensibilité** (sensitivity), le taux de vrais positifs, c'est-à-dire la proportion d'exemples positifs correctement identifiés.

Rappel ou Sensibilité

$$\frac{VP}{VP + FN}$$

Estimé / Réel	+	-
+	*VP*	*FN*
-	*FP*	*VN*

- Il est cependant très facile d'avoir un bon rappel en prédisant que tous les exemples sont positifs.
- Ainsi, ce critère ne peut pas être utilisé seul. On lui adjoint ainsi souvent la précision.

Matrice de confusion--Indicateurs de performances

- On appelle précision, ou valeur positive prédictive (positive predictive value, PPV) la proportion de prédictions correctes parmi les prédictions positives :

$$\textit{Précision} \quad \frac{VP}{VP + FP}$$

Estimé / Réel	+	-
+	*VP*	*FN*
-	*FP*	*VN*

- On appelle spécificité le taux de vrais négatifs, autrement dit la proportion d'exemples négatifs correctement identifiés comme tels.

$$\textit{Spécificité} \quad \frac{VN}{VN + FP}$$

Matrice de confusion--Indicateurs de performances

- Pour résumer rappel et précision en un seul nombre, on calculera la F-mesure (F-score ou F1-score), c'est la moyenne harmonique de la précision et du rappel:

$$F\text{-}measure \quad \frac{2 \times rappel \times précision}{rappel + précision} = \frac{2\,VP}{2\,VP + FP + FN}$$

Autres mesures:

$$FN\text{-}rate \quad \frac{FN}{VP + FN}$$

FN Rate « false negatif » : nombre de fois que la classe non prédite par le classifieur correspond à la vraie classe

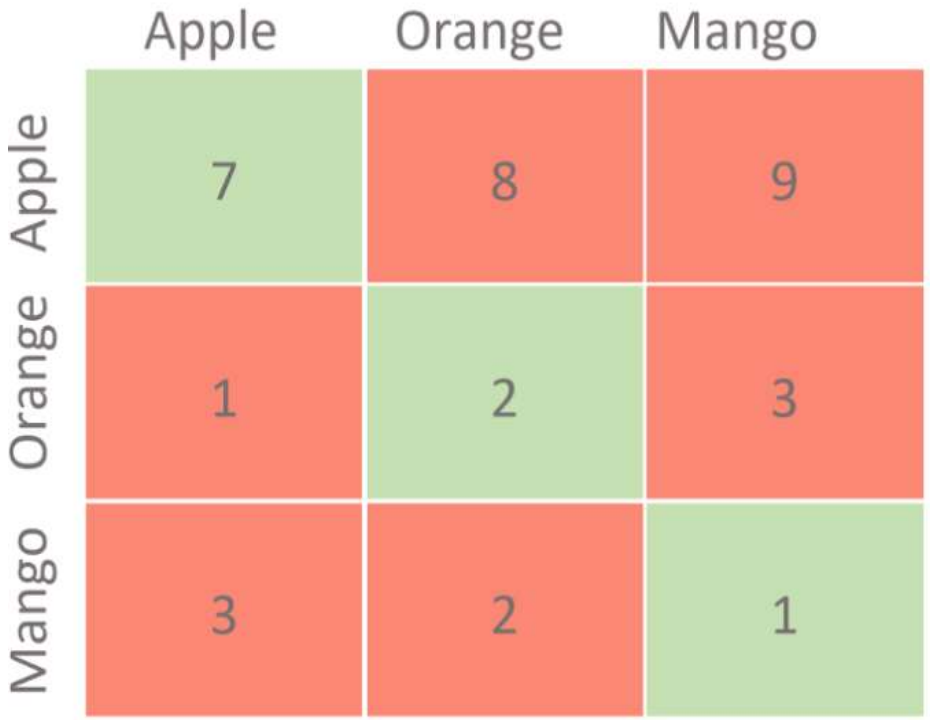
$$FP\text{-}rate \quad \frac{FP}{FP + VN}$$

FP Rate « false positive » : nombre de fois que la classe prédite par le classifieur ne correspond pas à la vraie classe.

Exemples de MC

		Predicted condition	
	Total 8 + 4 = 12	Cancer	Non-cancer
Actual condition	Cancer	6	2
	Non-cancer	1	3

Confusion Matrix for binary Classification



Confusion Matrix for Multi-Class Classification

Évaluation de méthodes de classification binaire retournant un score

- Si les algorithmes de classification ne retournent pas directement une étiquette de classe,
  - utilisent **une fonction de décision** qui doit ensuite être seuillée pour devenir une étiquette.
- Cette fonction de décision peut être un score arbitraire
  - par exemple, la proportion d'exemples positifs parmi les **k** plus proches voisins du point à étiqueter
  - ou la probabilité d'appartenir à la classe positive
- Plusieurs critères permettent d'évaluer la qualité de la fonction de décision avant seuillage.(exp: Courbe ROC)

Courbe ROC

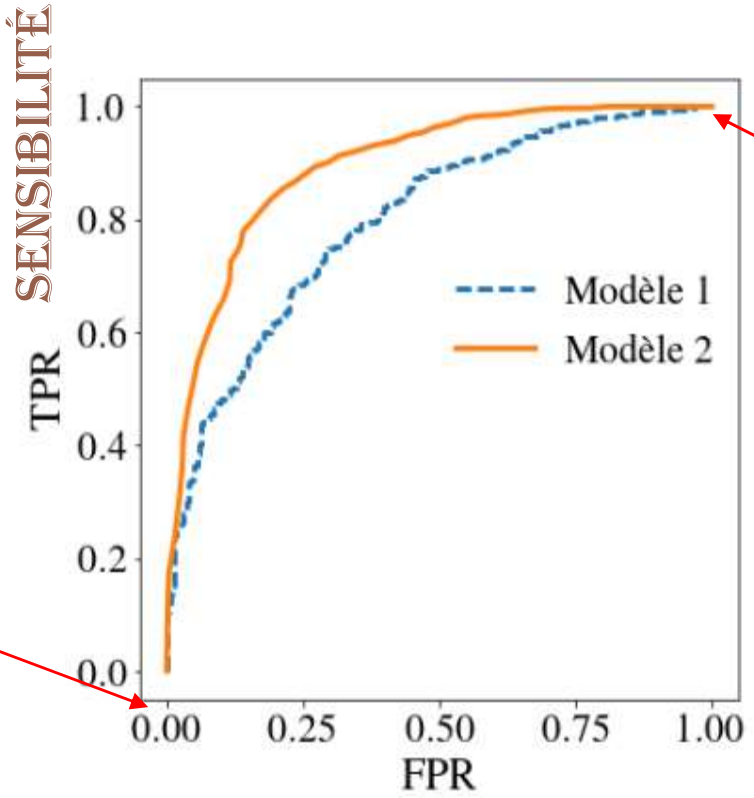
ROC (Receiver-Operator Characteristic)

- Le terme vient des télécommunications, où ces courbes servent à étudier si un système arrive à séparer le signal du bruit de fond.

- Décrive l'évolution de la sensibilité (TVP) en fonction du complémentaire à 1 de la spécificité, (1- spécificité ou TFP)
  - TVP = Rappel = Sensibilité = VP/Positifs
  - TFP = 1 – Spécificité = FP/Négatifs
- L'idée de la courbe ROC est de faire varier le « seuil » de 1 à 0 et, pour chaque cas, calculer le TVP et le TFP que l'on reporte dans un graphique

Courbe ROC

Le point(0, 0) apparaît quand on utilise comme seuil un nombre supérieur à la plus grande valeur retournée par la fonction de décision :
→ tous les exemples sont étiquetés négatifs.



le point (1, 1) apparaît quand on utilise pour seuil une valeur inférieure au plus petit score retourné par la fonction de décision :
→ tous les exemples sont alors étiquetés positif

Les courbes ROC de deux modèles

Courbe ROC

Construction de la courbe ROC (1/2)

Taux de vrais positifs Taux des Vrais Négatifs

Classer les données
selon un score décroissant

Individu	Score (+)	Classe
1	1	+
2	0.95	+
3	0.9	+
4	0.85	-
5	0.8	+
6	0.75	-
7	0.7	-
8	0.65	+
9	0.6	-
10	0.55	-
11	0.5	-
12	0.45	+
13	0.4	-
14	0.35	-
15	0.3	-
16	0.25	-
17	0.2	-
18	0.15	-
19	0.1	-
20	0.05	-

Positifs = 6
Négatifs = 14

Seuil = 1

	^positif	^négatif	Total
positf	1	5	6
négatif	0	14	14
Total	1	19	20

TVP = 1/6 = 0.2 ; TFP = 0/14 = 0

Seuil = 0.95

	^positif	^négatif	Total
positf	2	4	6
négatif	0	14	14
Total	2	18	20

TVP = 2/6 = 0.33 ; TFP = 0/14 = 0

Seuil = 0.9

	^positif	^négatif	Total
positf	3	3	6
négatif	0	14	14
Total	3	17	20

TVP = 3/6 = 0.5 ; TFP = 0/14 = 0

Seuil = 0.85

	^positif	^négatif	Total
positf	3	3	6
négatif	1	13	14
Total	4	16	20

TVP = 3/6 = 0.5 ; TFP = 1/14 = 0.07

Seuil = 0

	^positif	^négatif	Total
positf	6	0	6
négatif	14	0	14
Total	20	0	20

TVP = 6/6 = 1 ; TFP = 14/14 = 1

Courbe ROC

Construction de la courbe ROC (2/2)

Mettre en relation
TFP (abscisse) et TVP (ordonnée)

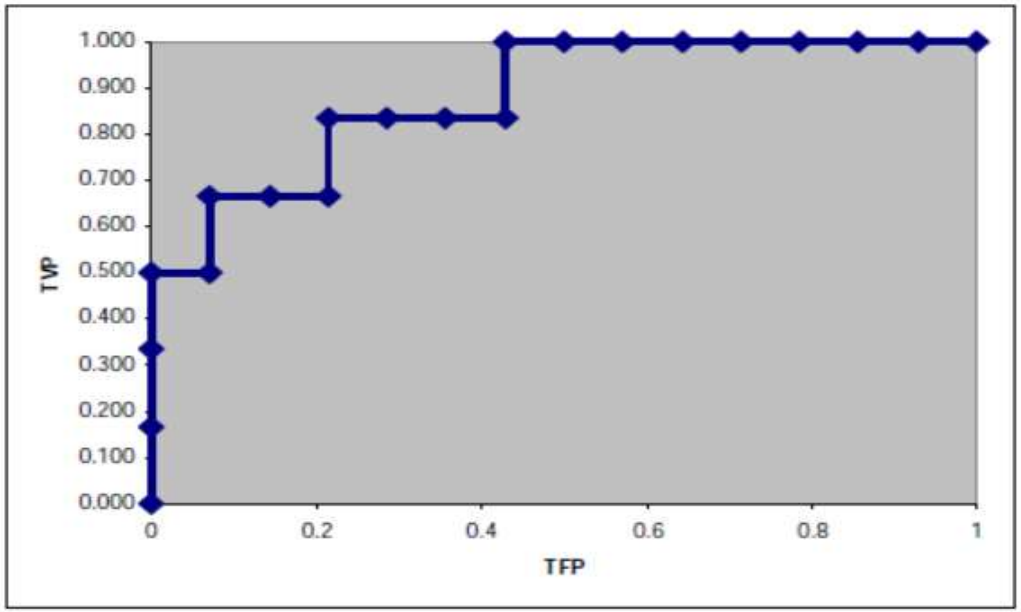
Individu	Score (+)	Classe	TFP	TVP
			0	0.000
1	1	+	0.000	0.167
2	0.95	+	0.000	0.333
3	0.9	+	0.000	0.500
4	0.85	-	0.071	0.500
5	0.8	+	0.071	0.667
6	0.75	-	0.143	0.667
7	0.7	-	0.214	0.667
8	0.65	+	0.214	0.833
9	0.6	-	0.286	0.833
10	0.55	-	0.357	0.833
11	0.5	-	0.429	0.833
12	0.45	+	0.429	1.000
13	0.4	-	0.500	1.000
14	0.35	-	0.571	1.000
15	0.3	-	0.643	1.000
16	0.25	-	0.714	1.000
17	0.2	-	0.786	1.000
18	0.15	-	0.857	1.000
19	0.1	-	0.929	1.000
20	0.05	-	1.000	1.000

Calcul pratique

TFP (i) = Nombre de négatifs parmi les « i » premiers / (nombre total des négatifs)

TVP (i) = Nombre de positifs parmi les « i » premiers / (nombre total des positifs)

Courbe ROC



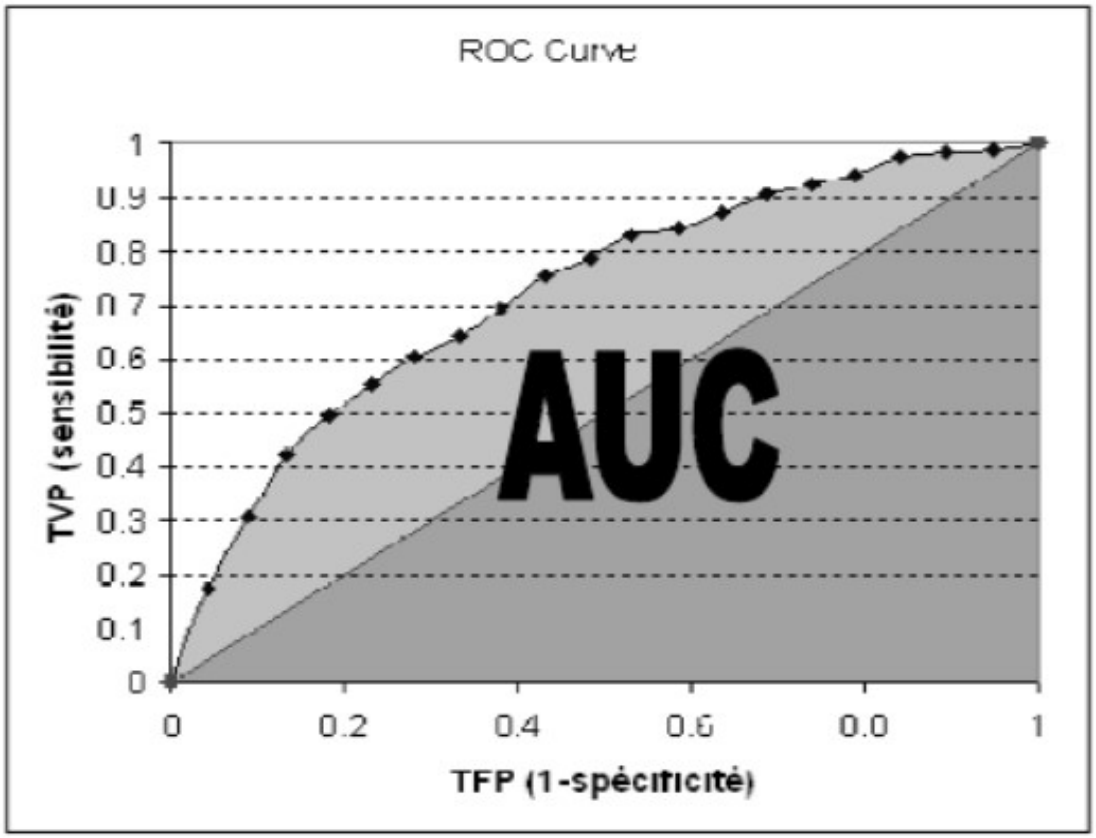
Courbe ROC et AUC

- L'aire sous la courbe (ou *Area Under the Curve – AUC*) est un indice synthétique calculé pour les courbes ROC.

- L'**AUC** correspond à la probabilité pour qu'un événement positif soit classé comme positif par le test sur l'étendue des valeurs seuil possibles.

- AUC est utilisé pour comparer les modèles !

Courbe ROC et AUC



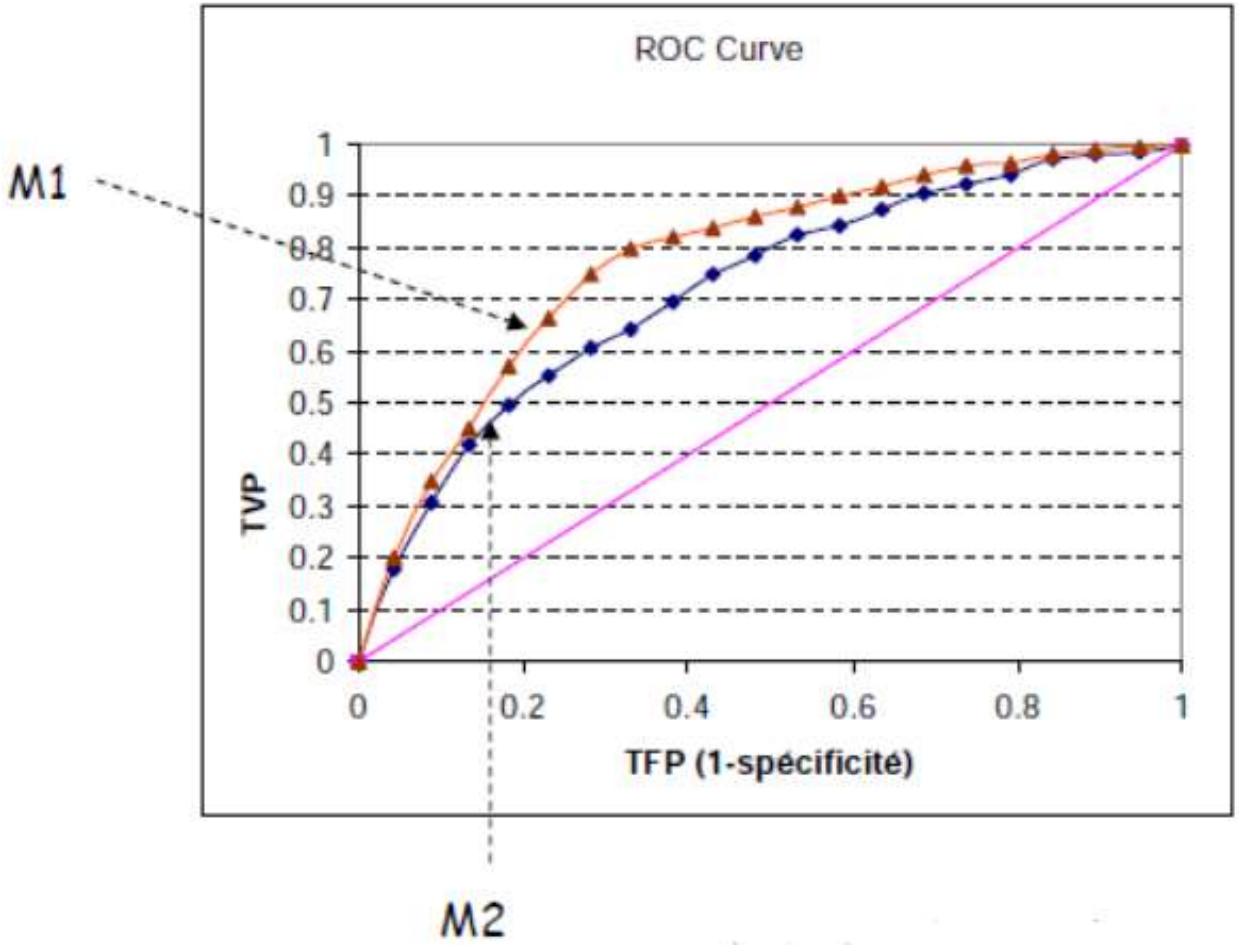
- Pour un modèle idéal, AUC=1 ,
- pour un modèle aléatoire, AUC=0.5 ;

(Symbolisée par la diagonale principale dans le graphique)

- Habituellement le modèle est bon dès lors que la valeur de l'AUC est supérieure à 0.7.
  - Un modèle bien discriminant doit avoir une AUC entre 0.87 et 0.9.
  - Un modèle ayant une AUC supérieure à 0.9 est excellent.

Courbe ROC--Interprétation : dominance

Question: Comment montrer que M1 sera toujours meilleur que M2, quelle que soit la matrice de coût de mauvaise affectation utilisée ?

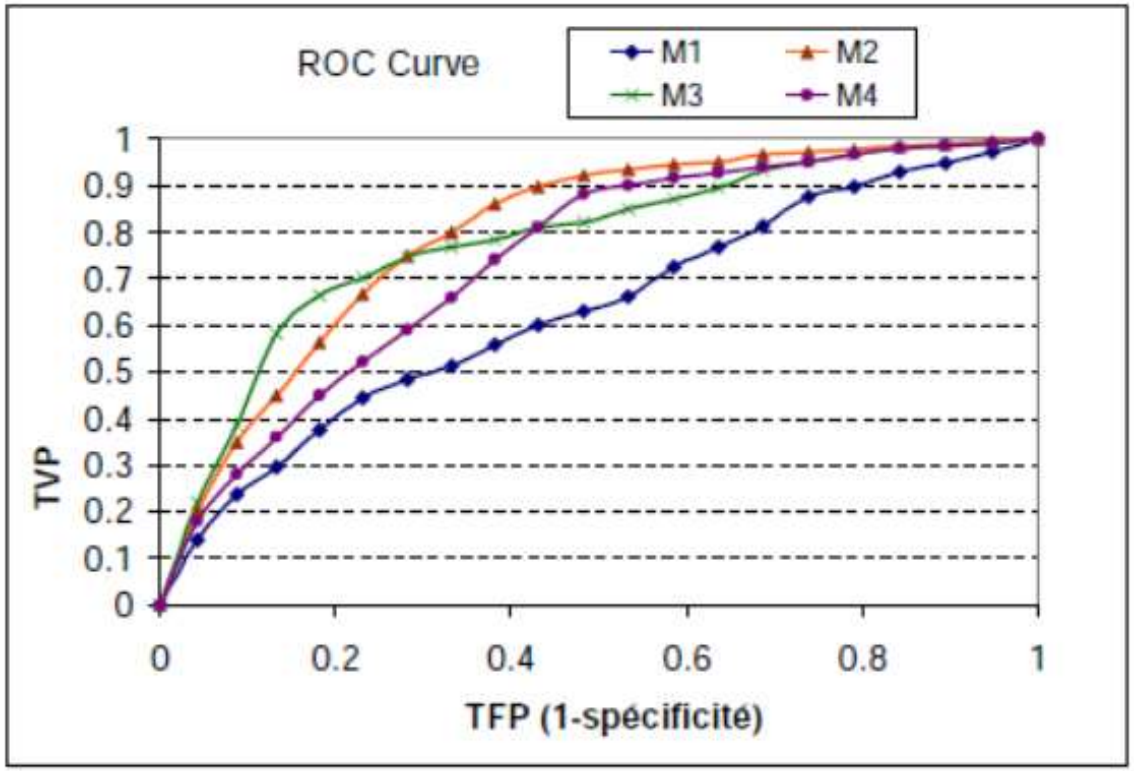


Interprétation : La courbe de M1 est toujours « au-dessus » de celle de M2 : il ne peut pas exister de situation (matrice de coût de mauvais classement) où M2 serait un meilleur modèle de prédiction

Courbe ROC--Enveloppe convexe : sélection de modèles

Question: Parmi un ensemble de modèles candidats, comment éliminer d'office ceux qui ne seront pas intéressants ?

- Enveloppe convexe: est formée par les courbes qui, à un moment ou à un autre, n'ont aucune courbe « au-dessus » d'elles.
- Les courbes situées sur cette enveloppe correspondent aux modèles qui sont potentiellement les plus performantes pour une matrice de coût donnée.
- Les modèles qui ne participent jamais à cette enveloppe peuvent être éliminés

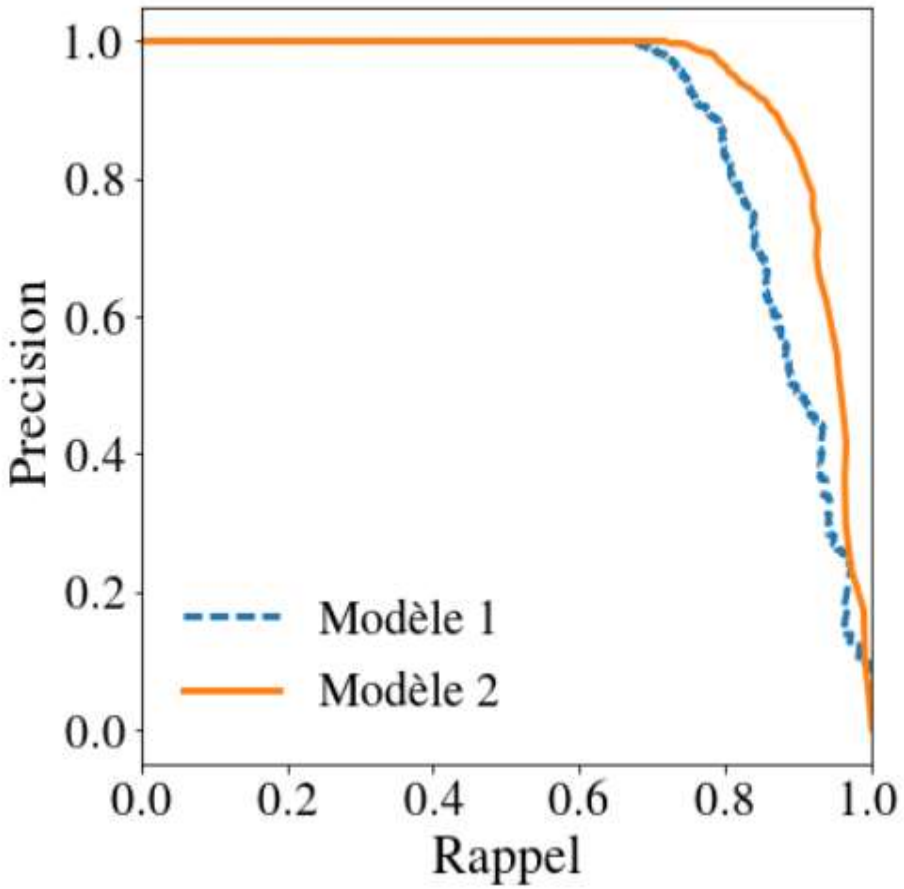


Interprétation:

- l'enveloppe convexe est formée par les courbes de M3 et M2.
- M1 est dominé par tous les modèles → il peut être éliminé
- M4 peut être meilleur que M3 dans certains cas, mais dans ces cas là, il sera moins bon que M2 → M4 peut être éliminé.

Courbe précision-rappel

- La courbe précision-rappel vient souvent complémenter la courbe ROC.
- On appelle courbe précision-rappel, ou Precision-Recall curve en anglais, la courbe décrivant l'évolution de la **précision** en fonction du **rappel**, lorsque le seuil de décision change.



Les courbes précision-rappel de deux modèles

Conclusion-ROC

- Dans de nombreuses applications, la courbe ROC fournit des informations plus intéressantes sur la qualité de l'apprentissage que le simple taux d'erreur.

  \+ C'est surtout vrai lorsque les classes sont très déséquilibrées, et lorsque le coût de mauvaise affectation est susceptible de modifications.

  \- Il faut néanmoins que l'on ait une classe cible (positive) clairement identifiée et que la méthode d'apprentissage puisse fournir un SCORE proportionnel à P(Y=+/X).

Erreurs de régression

- Dans le cas d'un problème de régression, le nombre d'erreurs n'est pas un critère approprié pour évaluer la performance.
- D'une part, à cause des imprécisions numériques, il est délicat de dire d'une prédiction à valeur réelle si elle est correcte ou non.

Question: quel est le meilleur modèle M1 ou M2?

- M1: Un modèle dont 50% des prédictions sont correctes à 0.1% près et les 50% autres sont très éloignées des vraies valeurs!
- M2: Un modèle modèle qui n'est correct qu'à 1% près, mais pour 100% des exemples ?

Solution: quantifier la performance d'un modèle de régression en fonction de l'écart entre les prédictions et les valeurs réelles.

Erreurs de régression

(Erreur quadratique moyenne (MSE)) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f(\vec{x}^{\,1}), f(\vec{x}^{\,2}), \ldots, f(\vec{x}^{\,n})$, on appelle *erreur quadratique moyenne*, ou *MSE* de l'anglais *mean squared error* la valeur

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(f(\vec{x}^{\,i}) - y^i\right)^2.$$

- Pour mesurer l'erreur **dans la même unité** que la cible, on lui préfère souvent sa **racine**:

RMSE) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f(\vec{x}^{\,1}), f(\vec{x}^{\,2}), \ldots, f(\vec{x}^{\,n})$, on appelle *racine de l'erreur quadratique moyenne*, ou *RMSE* de l'anglais *root mean squared error* la valeur

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(f(\vec{x}^{\,i}) - y^i\right)^2}.$$

- Dans le cas où les valeurs cibles couvrent **plusieurs ordres de grandeur**, on préfère parfois passer au **log** avant de comparer les prédictions aux vals réelles , afin de ne pas donner plus d'importance aux erreurs faites pour des valeurs plus élevées.

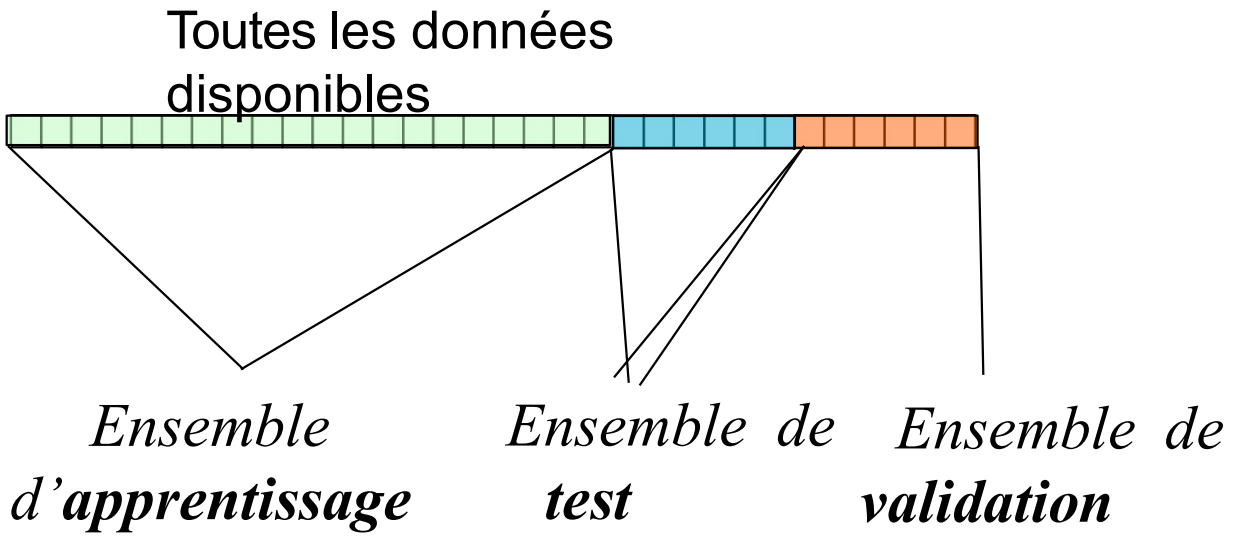
(RMSLE) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f(\vec{x}^{\,1}), f(\vec{x}^{\,2}), \ldots, f(\vec{x}^{\,n})$, on appelle *racine du log de l'erreur quadratique moyenne*, ou *RMSLE* de l'anglais *root mean squared log error* la valeur

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(f(\vec{x}^{\,i}) + 1\right) - \log\left(y^i + 1\right)\right)^2}.$$

Erreurs de régression--Normalisation

- L'interprétation de ces erreurs requiert néanmoins de connaître la distribution des valeurs cibles;
  - → *une RMSE de 1 cm n'aura pas la même signification selon qu'on essaie de prédire la taille d'humains ou celle de drosophiles (insecte).*
- Pour répondre à cela, il est possible de normaliser la somme des carrés des résidus non pas en en faisant la moyenne, mais en la comparant à la somme des distances des valeurs cibles à leur moyenne.

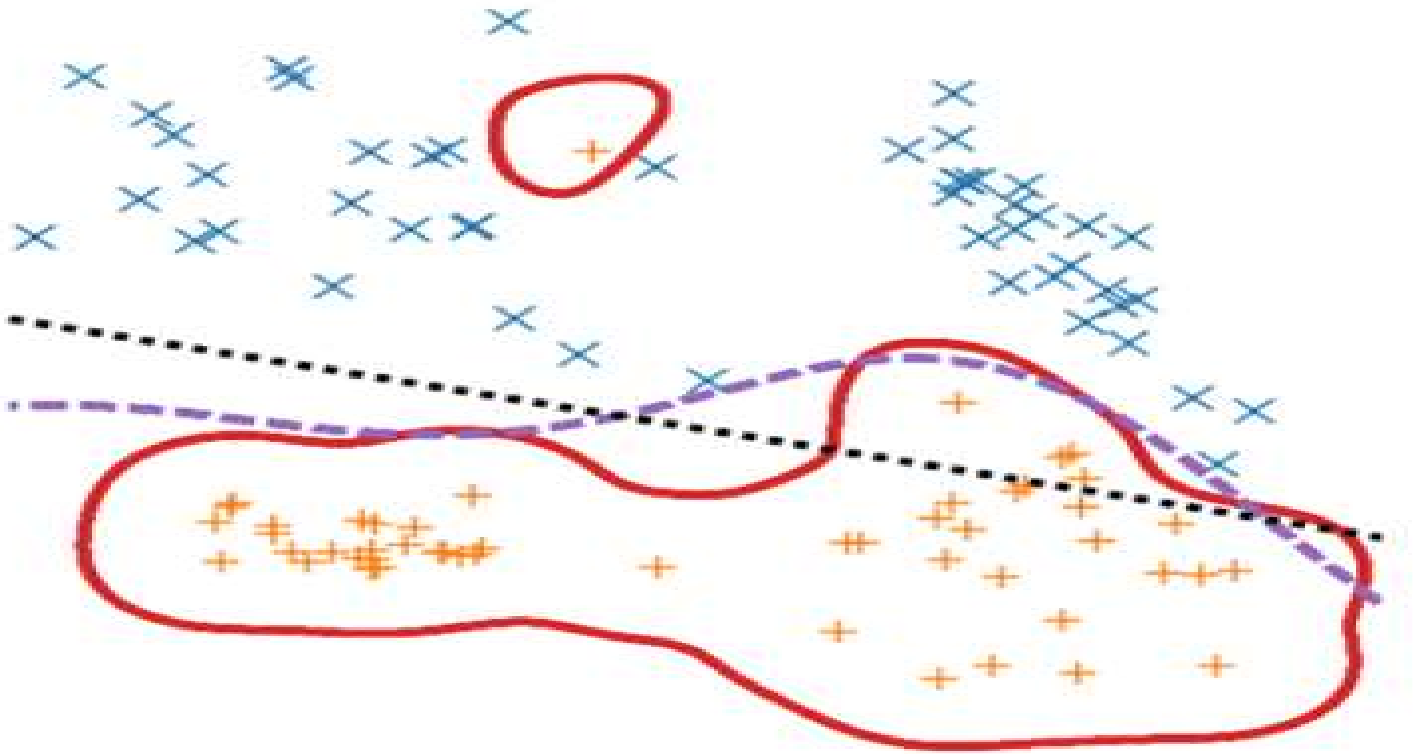
Ensembles de données (collections)



Généralisation et sur-apprentissage et l'ensembles de données

- **Généralisation:** c'est la capacité d'un modèle à faire des prédictions correctes sur de nouvelles données, qui n'ont pas été utilisées pour le construire.
- **Sur-apprentissage ou overfitting** : un modèle qui, plutôt que de capturer la nature des objets à étiqueter, modélise aussi le bruit et ne sera pas en mesure de généraliser qu'il sur-apprend.
- **Sous-apprentissage ou underfitting:** un modèle qui est trop simple pour avoir de bonnes performances même sur les données utilisées pour le construire qu'il sous-apprend.

Généralisation et sur-apprentissage et l'ensembles de données—Interprétation Exemple1

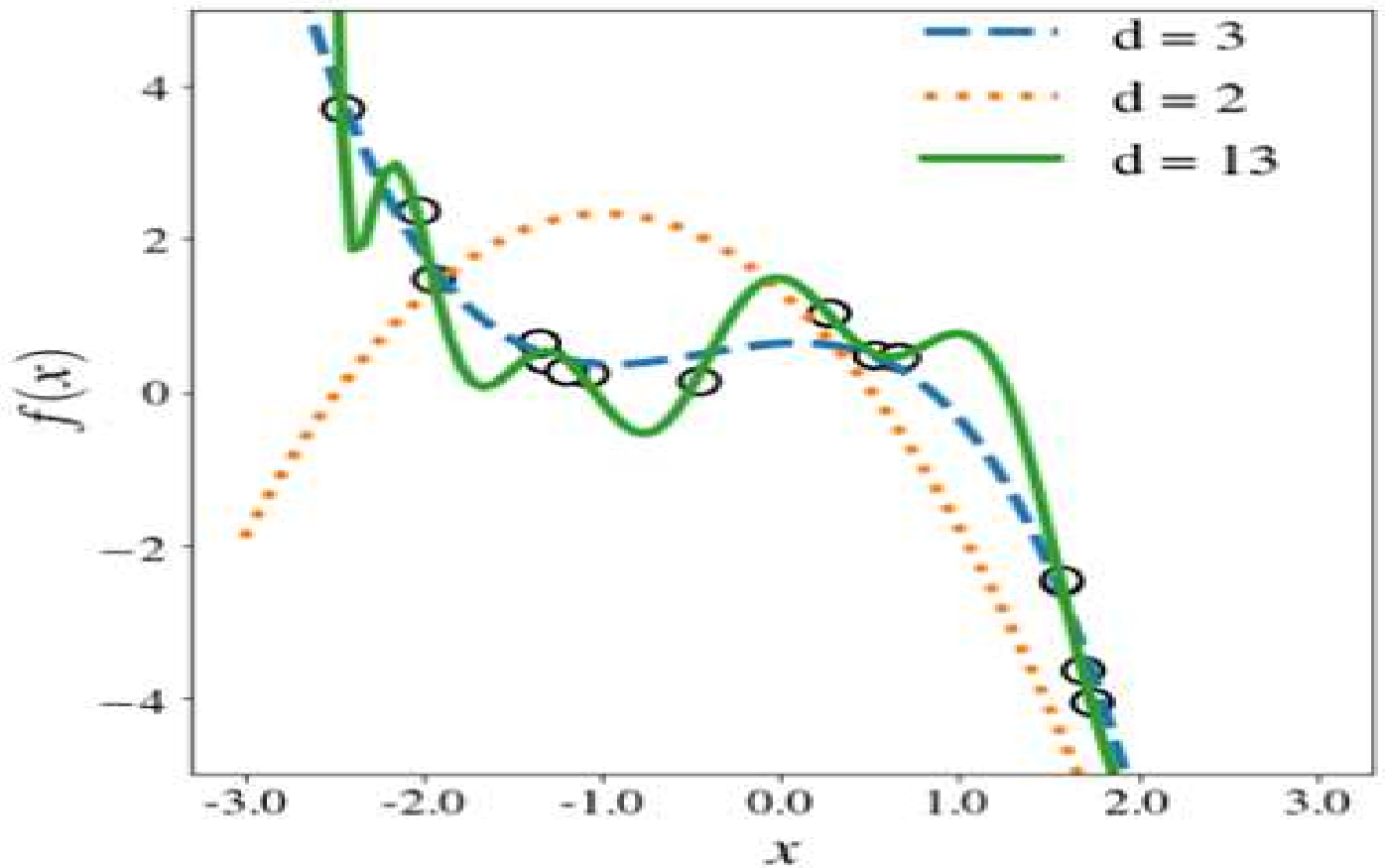


Pour séparer les observations négatives (x) des observations positives (+), la droite pointillée **sousapprend**.

→ La frontière de séparation en trait plein (rouge) ne fait aucune erreur sur les données mais est susceptible de **sur-apprendre.**

→ La frontière de séparation en trait discontinu est un bon compromis.

Généralisation et sur-apprentissage et l'ensembles de données—Interprétation Exemple2



Les étiquettes y des observations (représentées par des points) ont été générées à partir d'un polynôme de degré d = 3.

→Le modèle de degré d = 2 approxime très mal les données et **sous-apprend,** tandis que celui de degré d = 13, dont le risque empirique est plus faible, **surapprend**.

Points clefs

— Pour éviter le sur-apprentissage, il est essentiel lors de l'étape de sélection du modèle de valider les différents modèles testés sur un jeu de données différent de celui utilisé pour l'entraînement.

— Pour estimer la performance en généralisation d'un modèle, il est essentiel de l'évaluer sur des données qui n'ont été utilisées ni pour l'entraînement, ni pour la sélection de ce modèle.

— De nombreux critères permettent d'évaluer la performance prédictive d'un modèle. On les choisira en fonction de l'application.

— Pour interpréter la performance d'un modèle, il peut être utile de le comparer à une approche naïve.

Exemple

- Soit l'exemple d'un test clinique
- Il ne s'agit pas ici d'un modèle d'apprentissage automatique, mais d'un frottis de dépistage du cancer du col de l'utérus : il s'agit d'un examen beaucoup plus simple et moins invasif qu'un examen histologique, qui doit être interprété par un expert, et servira de vérité terrain.
- Les résultats d'une expérience menée sur 4 000 femmes âgées de 40 ans et plus sont présentés sur le tableau.

	Cancer	Pas de cancer	Total
Frottis +	190	210	400
Frottis -	10	3590	3600
Total	200	3800	4000

Interprétation

Le **rappel** est de **95%,** la **spécificité** de **94.5%,** mais la **précision** ne vaut que **47.5%**.

- Ce test est un bon outil de dépistage : la probabilité de n'avoir effectivement pas de cancer quand le frottis est négatif est élevée (3590/3600 ≈ 99.7%).
- MAIS, **c'est un mauvais outil diagnostique**, au sens où la probabilité de fausse alarme est très élevée!

Courbe ROC --Exemple

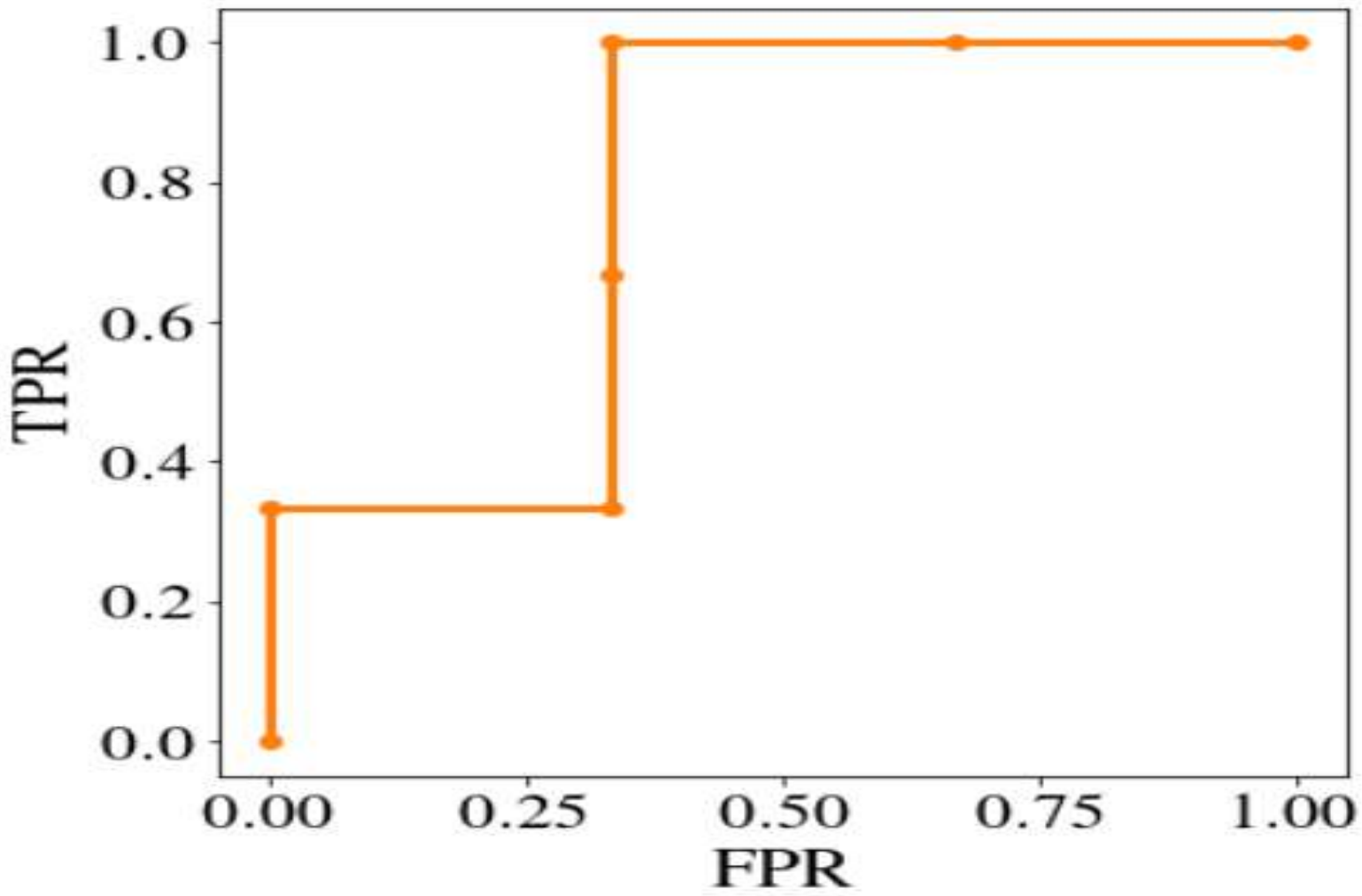
- Exemple de résultats d'une expérience de classification binaire, évaluée sur 6 échantillons.

Étiquette	+	-	+	+	-	-
Score	0.9	0.8	0.6	0.4	0.3	0.1

- Pour un seuil supérieur à 0.9, les 6 exemples sont étiquetés négatifs.
- On commence donc par le point (0, 0).
- Pour un seuil entre 0.95 et 0.9, seule la première observation est étiquetée positive.
- La sensibilité est donc de 1/3 tandis que l'antispécificité reste nulle.
- On peut continuer ainsi jusqu'à utiliser un seuil inférieur à 0.1 :

Seuil	> 0.9	0.8–0.9	0.6–0.8	0.4–0.6	0.3–0.4	0.1–0.3	< 0.1
TP/P	0	1/3	1/3	2/3	1	1	1
FP/P	0	0	1/3	1/3	1/3	2/3	1

Courbe ROC –Exemple (suite)



Courbe précision-rappel--Exemple

Étiquette	+	-	+	+	-	-
Score	0.9	0.8	0.6	0.4	0.3	0.1

- Les valeurs de la précision et du rappel sont les suivantes:

Seuil	> 0.9	0.8–0.9	0.6–0.8	0.4–0.6	0.3–0.4	0.1–0.3	< 0.1
Rappel	0	1/3	1/3	2/3	1	1	1
Précision	-	1	1/2	2/3	3/4	3/5	3/6

On obtient donc la courbe précision-rappel