

Introduction au machine learning

Séance 1 – Principes

Christophe Ambroise 

UEVE, UMR CNRS 8071

4 septembre 2026

1

Introduction

2

Ouvrages de référence



Figure 1

Les quatre sont disponibles gratuitement en PDF, légalement, sur les pages des auteurs.

3

Lequel pour quoi

- **Hastie, Tibshirani & Friedman**. Théorie de la décision, analyse discriminante, arbres, régularisation.
- **James, Witten, Hastie, Tibshirani & Taylor** — même matière, moins de mathématiques, avec code Python. Le bon point d'entrée en cas de difficulté.
- **Bishop** — angle probabiliste et bayésien. Traitement complet des mélanges et des méthodes à noyaux.
- **Murphy** — le plus large et le plus récent. Mon préféré.

4

Evaluation

- Un partiel de 2 heures, sur table, à la fin du semestre. Il couvre tout le cours
- Un projet facultatif de 2 à 3 semaines à rendre sur GitHub. Il consiste à appliquer les méthodes vues en cours à un jeu de données réel.

5

Le cours en résumé

- **Extraire** de l'information à partir de données
- **Apprendre** à **prédire** à partir d'exemples

6

Repères historiques

7

1943–1959

Année Jalon

- | | | |
|------|---------------------------|------------------------------------|
| 1943 | Neurone formel | — Warren McCulloch et Walter Pitts |
| 1950 | Jeu de l'imitation | — Alan Turing |
| 1956 | Intelligence artificielle | — atelier de Dartmouth |
| 1958 | Perceptron | — Frank Rosenblatt |
| 1959 | Apprentissage automatique | — jeu de dames d'Arthur Samuel |

8

1984–2017

Année Jalon

- | | | |
|------|------------------------------|---|
| 1984 | CART | — arbres de classification et de régression |
| 1986 | Rétropropagation | — Rumelhart, Hinton et Williams |
| 1995 | Machines à vecteurs supports | — Cortes et Vapnik |
| 2001 | Forêts aléatoires | — Leo Breiman |
| 2012 | AlexNet | — ImageNet et calcul sur processeur graphique |
| 2017 | Transformers | — mécanisme d'attention |

9

Pourquoi l'Intelligence Artificielle est si présente en 2026:

Trois conditions réunies depuis peu.

- Des **données** massives, produites par défaut (web, capteur, livres,...).
- Une puissance de **calcul** devenue abordable (GPU, centre de calcul,...)
- Des **algorithmes** capables d'en tirer parti (Réseaux de neurones profonds, forêt aléatoires, ...)

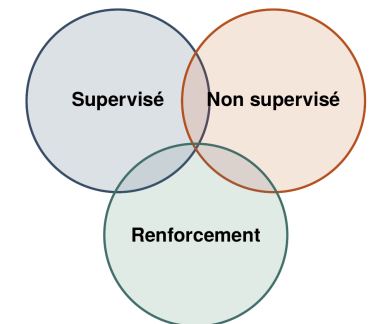
10

Trois familles de méthodes

- **Supervisé** — des données (x_i) étiquetées (y_i) . Classification, régression.

$$(X, Y) \sim p \quad \text{inconnue,} \\ D_n = \{(x_i, y_i)\}_{i=1}^n \quad \text{observé.}$$

- **Non supervisé** — des données sans étiquettes (x_i) . Recherche de groupes, réduction de dimension.
- **Par renforcement** — un agent, un environnement, une récompense.



11

Iris, le fil du cours

12

Un tableau

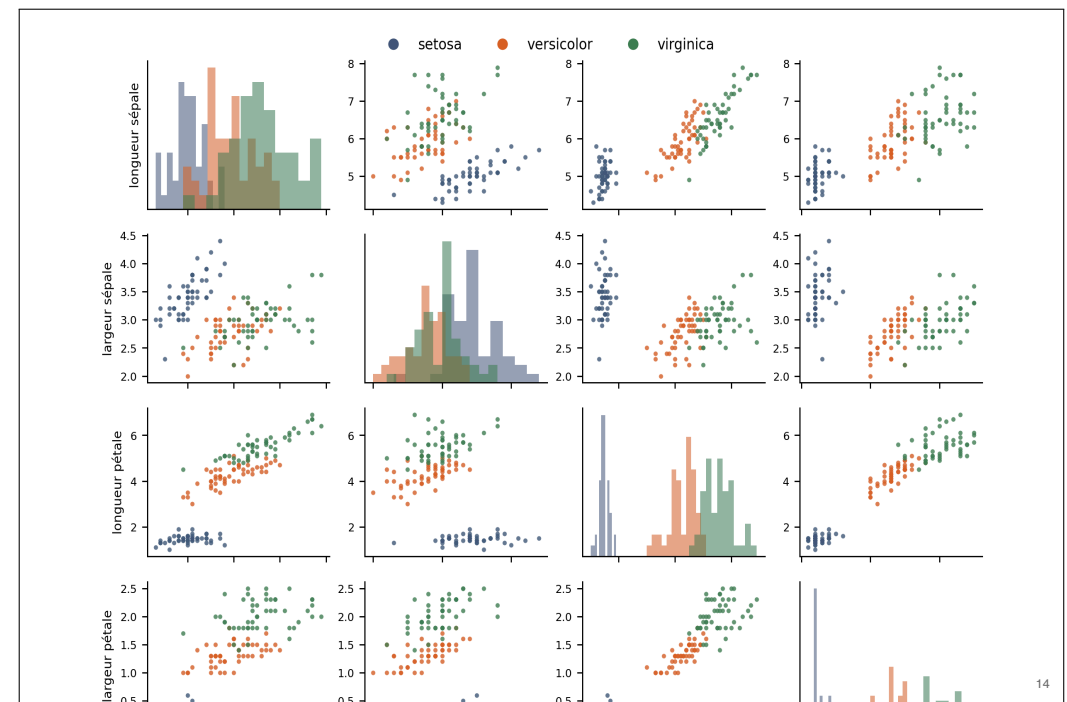
	longueur sépal	largeur sépal	longueur pétale	largeur pétale	espèce
1	5,1	3,5	1,4	0,2	setosa
51	7,0	3,2	4,7	1,4	versicolor
101	6,3	3,3	6,0	2,5	virginica

Une ligne : une observation. Une colonne : une variable.

Recueilli par Edgar Anderson, publié par Ronald Fisher en 1936.

13

Les quatre mesures



14

Premier aperçu

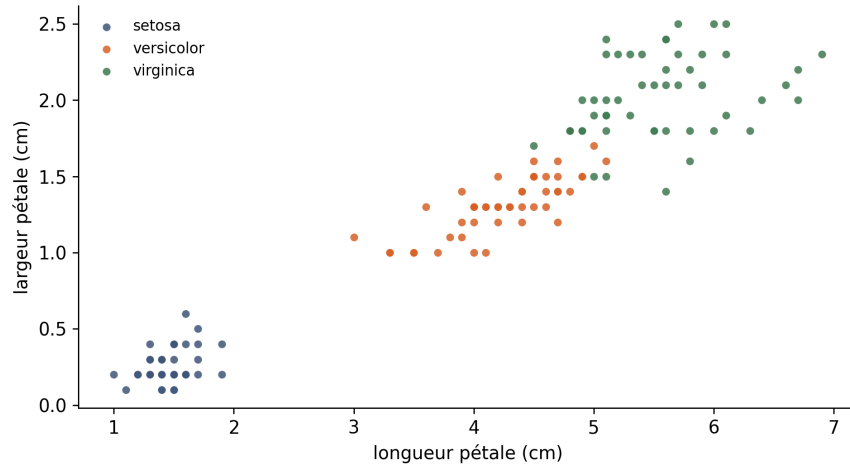


Figure 3

Deux variables suffisent à isoler *setosa*. Les deux autres espèces se chevauchent.

Deux usages du même tableau

Avec l'espèce

Prédire l'étiquette d'une fleur nouvelle.

Mêmes 150 fleurs, mêmes deux mesures. Deux questions différentes, deux figures différentes.

Sans l'espèce

Chercher des groupes dans le nuage.

Supervisé : prédire l'espèce

Une fleur mesurée, une espèce à annoncer. Les 150 étiquettes servent à construire la règle.

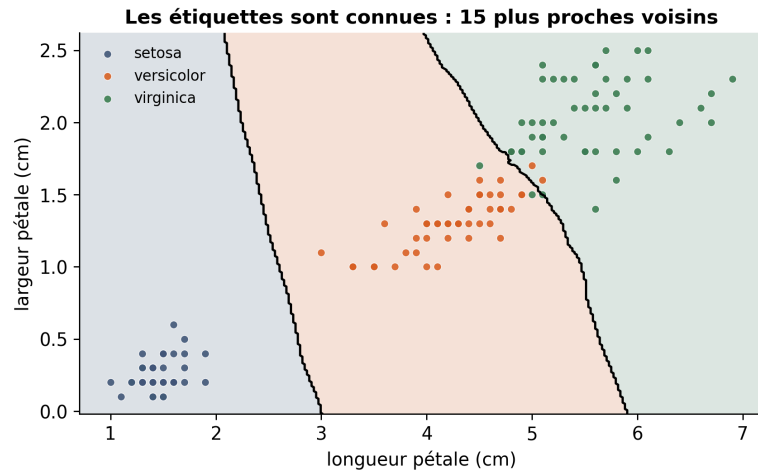


Figure 4

17

Frontières et apprentissage supervisé

Trois régions, deux frontières. Une fleur nouvelle tombe dans une région, et reçoit l'espèce correspondante.

Setosa se sépare d'un trait net. Entre *versicolor* et *virginica*, la frontière traverse une zone où les deux espèces se mélangent.

La règle est jugée sur des fleurs qu'elle n'a jamais vues.

18

Non supervisé : chercher des groupes

Aucune étiquette. Une seule consigne : trouver trois paquets compacts.

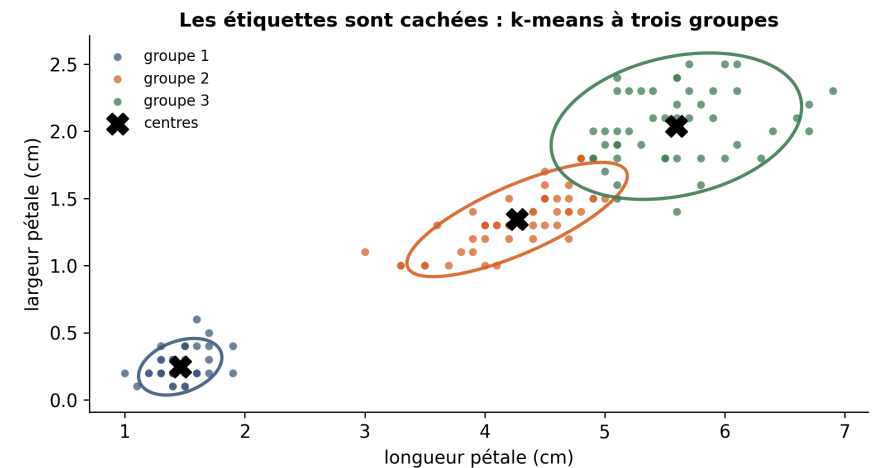


Figure 5

19

Le modèle de mélange implique des groupes

Chaque ellipse résume la dispersion d'un groupe à deux écarts-types. Les croix marquent les centres.

Rien dans ce calcul ne connaît le mot « espèce ». La structure vient de la géométrie du nuage, pas d'une réponse fournie.

20

Les groupes ne sont pas les espèces

Table 1

groupe	setosa	versicolor	virginica
1	50	0	0
2	0	48	4
3	0	2	46

Setosa est retrouvée exactement. Six fleurs sur cent cinquante changent de camp entre *versicolor* et *virginica*.

Un groupe est une réponse à une question de géométrie. Il ne coïncide avec une catégorie du monde que si celle-ci se

traduit en distances.

21

Qu'est-ce qu'apprendre ? Apprentissage supervisé

22

Apprendre à partir de données

- Un échantillon.
- Une famille de fonctions de décision.
- Un choix dans cette famille.

$$D_n = \{(x_i, y_i)\}_{i=1}^n, \quad \hat{f} \in \mathcal{F}.$$

Les observations sont *souvent* supposées indépendantes et de même loi.

Hypothèse forte: Invalidé par

- une dépendance temporelle ou
- un effet de groupe.

Le contexte peut évoluer et la loi changer

23

Estimer ou prédire

Estimer — comprendre

- Une quantité inconnue : θ
- Cible : la proximité entre $\hat{\theta}$ et θ .
- Incertitude, biais, variance.

Prédire — généraliser

- Une réponse nouvelle : Y_{new}
- Cible : l'erreur de $\hat{f}(X_{\text{new}})$.
- Performance hors échantillon d'apprentissage.

Deux objectifs distincts

- Estimer : retrouver une caractéristique du mécanisme.
- Prédire : être précis sur de nouvelles observations.

24

Décider sur quelle base ?

L'estimation de quantités inconnues est un problème central des statistiques.

Pour juger de la qualité d'une quantité estimée, la théorie de la décision propose de se donner **un critère**.

Fonction de coût

$L(\text{valeur réelle, décision})$

Elle mesure le prix payé pour une décision, si l'on connaissait la vérité.

Tout ce qui suit découle de ce choix.

25

Un détour par l'estimation

Problème. Estimer le paramètre θ d'une densité f , à partir de $D_n = (x_1, \dots, x_n)$, réalisations indépendantes de cette loi.

L'approche fréquentiste juge l'estimateur $\hat{\theta}$ par son **risque** :

$$R(\theta, \hat{\theta}) = \mathbb{E}[L(\theta, \hat{\theta}(D_n))] = \int_{\mathcal{X}} L(\theta, \hat{\theta}(D_n)) f(D_n | \theta) dD_n$$

Le risque est une **espérance** : une moyenne sur tous les échantillons qu'on aurait pu observer. Nous n'en observons qu'un.

26

Coût quadratique et première décomposition

Avec $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$:

$$\begin{aligned} R(\theta, \hat{\theta}) &= \mathbb{E}[(\theta - \hat{\theta})^2] \\ &= \underbrace{\text{Var}[\hat{\theta}]}_{\text{variance}} + \underbrace{(\mathbb{E}_{\theta}[\hat{\theta}] - \theta)^2}_{\text{biais}^2} \end{aligned}$$

Parmi les estimateurs **sans biais**, le meilleur est celui de variance minimale.

Cette décomposition distingue l'erreur systématique de la sensibilité de l'estimateur à l'échantillon observé.

27

Deux types de problèmes

Classification

- $Z = (X, Y) \in \mathcal{Z} = \mathbb{R}^d \times \{-1, 1\}$
- Objectif : pour une nouvelle observation x , estimer $P(Y | X = x)$

Régression

- $Z = (X, Y) \in \mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$
- Objectif : pour une nouvelle observation x , estimer $\mathbb{E}[Y | X = x]$

28

Apprendre, formellement

Apprendre consiste à chercher une **bonne** fonction dans un espace de fonctions $\mathcal{F}$.

Exemples

- Classification : $\hat{y} = f(x; \theta) = \text{sign}(\theta^\top x)$
- Régression : $\hat{y} = f(x; \theta) = \theta^\top x$

Le choix de $\mathcal{F}$ est la première décision, lourde de conséquences.

29

Le coût, concrètement

Fonction de coût $L : \mathcal{Z} \times \mathcal{F} \rightarrow \mathbb{R}_+$.

En classification — le coût 0-1 :

$$L(z, f) = L((x, y), f) = \mathbb{1}_{\{y \neq f(x)\}}$$

En régression — le coût quadratique :

$$L(z, f) = L((x, y), f) = (y - f(x))^2$$

Le coût 0-1 traite un faux négatif et un faux positif comme équivalents. Sur des données médicales, c'est un choix discutable.

30

Le risque, et son minimiseur

Le **risque attendu** d'une fonction f :

$$R(f) = \mathbb{E}_Z[L(z, f)] = \int L(z, f) p(z) dz$$

Principe d'induction

$$f^* = \arg \min_{f \in \mathcal{F}} R(f)$$

...

Ce minimum est le meilleur risque accessible dans la classe $\mathcal{F}$. Le **risque de Bayes** est le minimum obtenu sur l'ensemble de toutes

les règles de décision.

31

Sous le coût 0-1 : la règle MAP

Pour une observation x , le risque conditionnel associé à la décision k est

$$\begin{aligned} R(k | x) &= \mathbb{E}[\mathbb{1}_{\{Y \neq k\}} | X = x] \\ &= P(Y \neq k | X = x) \\ &= 1 - P(Y = k | X = x). \end{aligned}$$

$$\begin{aligned} g_B(x) &= \arg \min_k R(k | x) \\ &= \arg \max_k P(Y = k | X = x) \\ &= \arg \max_k \pi_k f_k(x). \end{aligned}$$

où $\pi_k = P(Y = k)$ et $f_k(x) = p(x | Y = k)$; le dénominateur $p(x)$ de la formule de Bayes ne dépend pas de k .

Sous le coût 0-1, le classifieur de Bayes est donc la règle du **maximum a posteriori** (MAP).

32

La règle de Bayes, en image

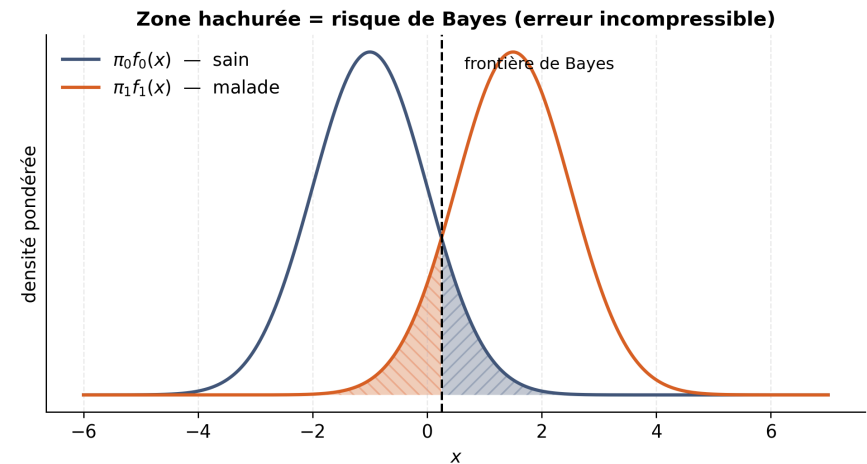


Figure 6

On décide « malade » là où $\pi_1 f_1(x) > \pi_0 f_0(x)$. La frontière est le point d'égalité. Dans le cas binaire,

$$R_B = R(g_B) = \int \min\{\pi_0 f_0(x), \pi_1 f_1(x)\} dx.$$

La zone hachurée représente ce risque minimal.

33

Le risque de Bayes n'est pas zéro

La zone hachurée ne disparaîtra jamais :

- deux individus ayant **le même x** peuvent avoir des y différents
- aucune fonction de x ne peut les séparer

Une erreur résiduelle n'est pas nécessairement un échec du modèle. Elle peut être une propriété du problème.

C'est le premier réflexe à acquérir : avant de vouloir descendre à 0 % d'erreur, demandez-vous si c'est possible.

34

Un coût, deux erreurs

Sur un dépistage, deux façons de se tromper :

	Prédit sain	Prédit malade
Réellement sain	✓	faux positif
Réellement malade	faux négatif	✓

- Un faux positif : un examen complémentaire inutile.
- Un faux négatif : un infarctus non prévenu.

Le coût 0-1 les compte pour un chacun, c'est un choix.

35

Coûts asymétriques

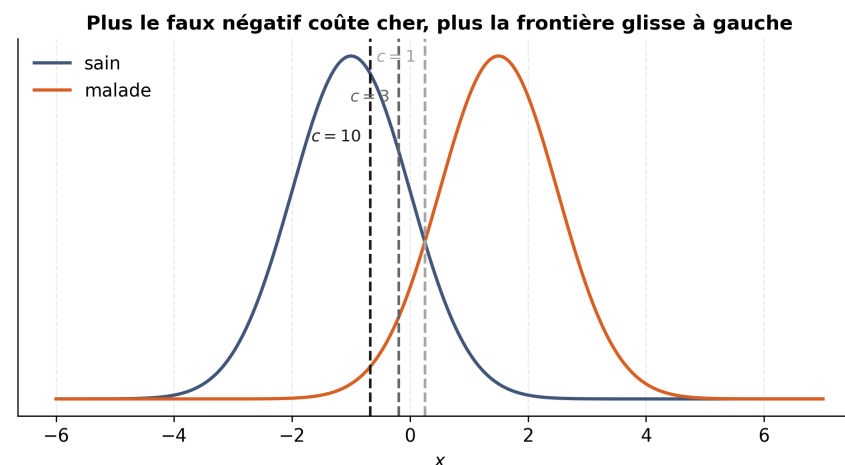


Figure 7

Si un faux négatif coûte c fois un faux positif, on décide « malade » dès que $c \pi_1 f_1(x) > \pi_0 f_0(x)$: la frontière glisse.

36

Ce que cela signifie

Changer les coûts change la décision.

Deux conséquences :

- Un classifieur encode un arbitrage.
- Les métriques d'équité formalisent, elles aussi, des choix et des arbitrages.

37

En bref

1. Apprendre = choisir $f \in \mathcal{F}$ minimisant un risque.
2. Le risque dépend d'un **coût**, qui est une décision, pas une donnée.
3. Le meilleur risque atteignable est le **risque de Bayes** — rarement nul.
4. $p(z)$ **est inconnue** : tout le reste en découle.

—Pause—

38

Loi inconnue

39

Remplacer l'espérance par une moyenne

Puisque $p(z)$ est inconnue, remplaçons l'intégrale par ce que nous avons :

$$\hat{R}(f, D_n) = \frac{1}{n} \sum_{i=1}^n L(z_i, f)$$

C'est le **risque empirique**.

Il est **sans biais** : $\mathbb{E}_D[\hat{R}(f, D)] = R(f)$.

Cette propriété est vraie **à f fixée**. Elle ne s'applique plus directement si les mêmes données servent à choisir f puis à

40

Minimisation du risque empirique

$$f^*(D_n) = \arg \min_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

C'est le principe qui se cache derrière la méthode `.fit()`.

Mais f^* **dépend de D_n** . Évaluer $f^*(D_n)$ sur D_n n'est plus une estimation sans biais.

41

Erreur d'apprentissage vs erreur de généralisation

Erreur de généralisation — ce qui nous intéresse, incalculable :

$$R(f^*(D_n)) = \mathbb{E}_z[L(z, f^*(D_n)) \mid D_n] = \int L(z, f^*(D_n)) p(z) dz$$

où $z \sim p$ est **indépendante de D_n** : une observation jamais vue. Dans l'intégrale, D_n est fixé — donc $R(f^*(D_n))$ reste **aléatoire en D_n** .

Erreur d'apprentissage — ce que nous calculons :

$$\hat{R}(f^*(D_n), D_n) = \frac{1}{n} \sum_{i=1}^n L(z_i, f^*(D_n))$$

Or, par définition de f^* :

42

L'erreur d'apprentissage est optimiste

$f^*(D_n)$ a été choisie **pour être bonne sur D_n** . Elle est donc meilleure sur D_n que sur tout autre échantillon $D_{n'}$ tiré de la même loi.

$$\mathbb{E}[R(f^*(D_n)) - \hat{R}(f^*(D_n), D_n)] \geq 0$$

Un taux d'erreur mesuré sur les données d'apprentissage n'est pas un résultat. C'est une mesure de mémorisation.

Dans un rapport, présenter l'erreur d'apprentissage comme une performance du modèle constitue une erreur méthodologique

43

Voir le phénomène

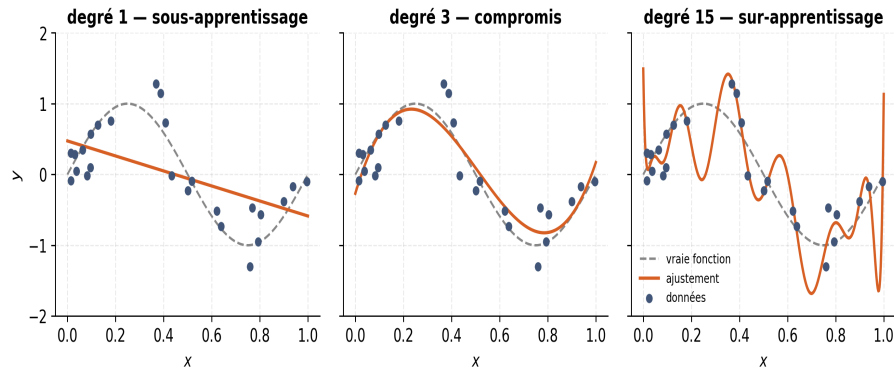


Figure 8

Même vraie fonction, mêmes 25 observations, trois espaces $\mathcal{F}$ de capacités croissantes. À droite, l'erreur d'apprentissage est beaucoup plus faible, mais la courbe s'écarte de la vraie fonction entre les observations.

De combien peut-on se tromper ?

Avec un coût 0-1 et $\mathcal{F}$ finie :

$$P\left(\max_{f \in \mathcal{F}} |R(f) - \hat{R}(f, D_n)| > \epsilon\right) \leq 2 |\mathcal{F}| e^{-2n\epsilon^2}$$

Deux leviers, et un seul est gratuit :

- n grand $\rightarrow$ la borne s'effondre exponentiellement
- $|\mathcal{F}|$ grand $\rightarrow$ la borne se dégrade

Et si $\mathcal{F}$ est infinie ?

$|\mathcal{F}|$ n'a plus de sens : on lui substitue la **capacité**, ou dimension de Vapnik-Chervonenkis.

En classification, la VC-dimension est le plus grand n tel qu'il existe un ensemble de n points que $\mathcal{F}$ peut séparer **quel que soit l'étiquetage**.

On dit alors que $\mathcal{F}$ pulvérise ces points.

VC-dimension d'un classifieur linéaire dans $\mathbb{R}^2$



Trois points, oui, quel que soit l'étiquetage. Quatre points, non.

La VC-dimension d'une droite dans le plan vaut $3 = d + 1$.

47

Capacité contre erreur

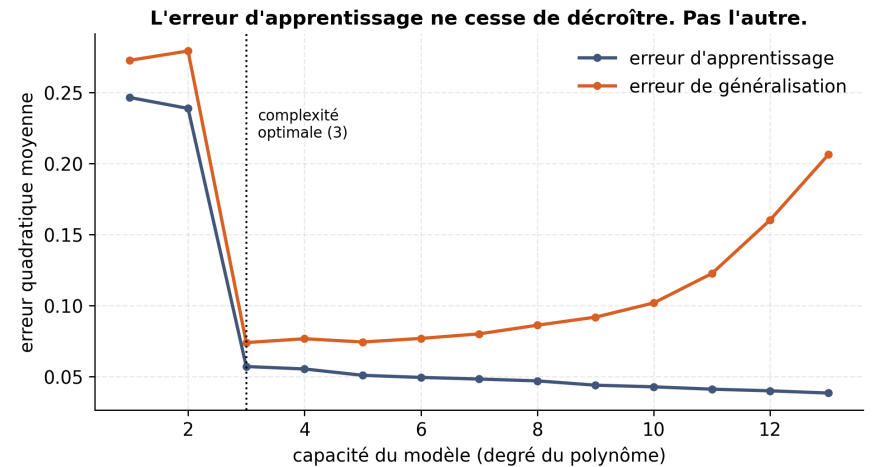


Figure 9

La courbe est obtenue en moyennant les erreurs sur 200 échantillons.

48

Le risque quadratique, en détail

Supposons $Z = (X, Y)$ et $L(Y, f(X)) = (Y - f(X))^2$. Le risque

$$R(f) = \mathbb{E}_{X,Y} [(Y - f(X))^2]$$

est minimisé par $f(x) = \mathbb{E}[Y \mid X = x]$.

Écrivons $Y = f(X) + \epsilon$, avec $\mathbb{E}[\epsilon] = 0$, et soit $\hat{f} = \hat{f}(D_n)$ construite sur l'échantillon.

49

La décomposition, en trois morceaux

En conditionnant par rapport à x :

$$\mathbb{E}_{D_n, Y|x}[(Y - \hat{f}(x))^2 | x] = \underbrace{\mathbb{E}_{D_n, Y|x}[(Y - f(x))^2 | x]}_{\sigma^2} + \mathbb{E}_{D_n, Y|x}[(f(x) - \hat{f}(x))^2 | x]$$

et le second terme se scinde encore :

$$\mathbb{E}_{D_n, Y|x}[(f(x) - \hat{f}(x))^2 | x] = \text{Biais}^2(\hat{f}(x)) + \text{Var}(\hat{f}(x))$$

$$\text{erreur} = \underbrace{\text{biais}^2}_{\mathcal{F} \text{ trop pauvre}} + \underbrace{\text{variance}}_{\text{sensibilité à } D_n} + \underbrace{\sigma^2}_{\text{bruit}}$$

50

Les trois termes, en mots

- **Biais** — $\mathcal{F}$ ne contient pas la solution optimale. Le modèle est trop pauvre pour le problème.
- **Variance** — avec un autre échantillon $D_{n'}$ tiré de la même loi, on aurait obtenu une fonction différente.
- **Bruit** — même la solution optimale se trompe : pour un même x , plusieurs y sont possibles.

Seuls les deux premiers dépendent de vous.

51

Biais et variance en régression non linéaire

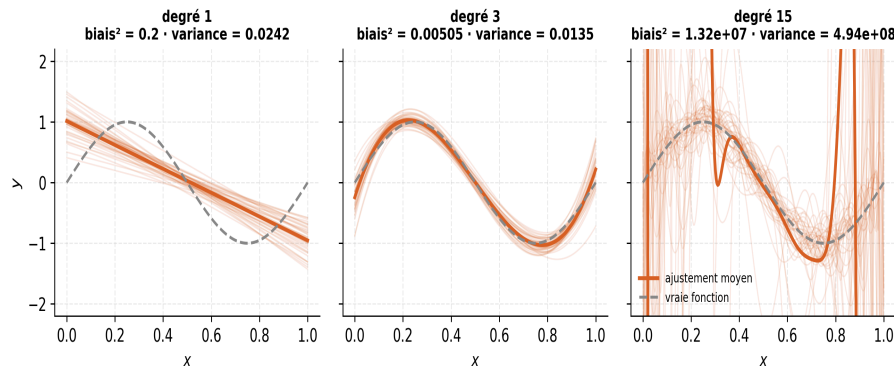


Figure 10

40 échantillons différents, trois capacités. À gauche : toutes les courbes se ressemblent, mais aucune ne colle. À droite : elles collent, mais chacune autrement.

52

Le compromis est forcé

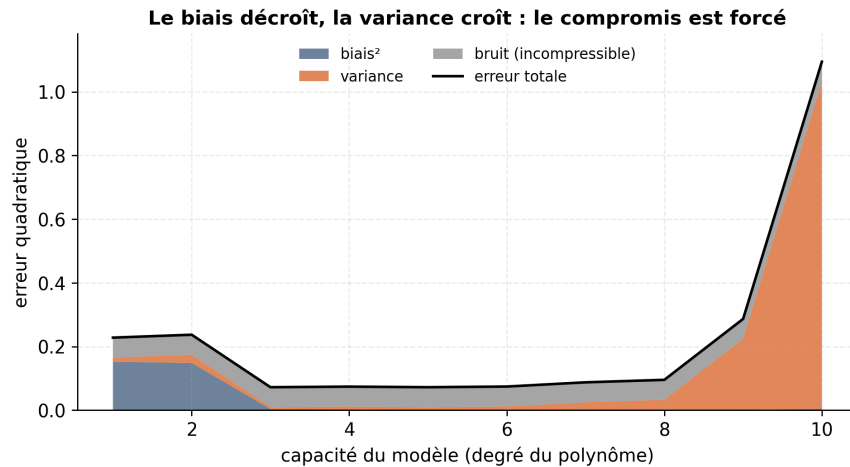


Figure 11

Quand la capacité de $\mathcal{F}$ croît, le biais diminue **et** la variance augmente. On ne peut pas gagner sur les deux tableaux.

53

Sur-apprentissage et sous-apprentissage

- Capacité **trop grande** → **sur-apprentissage** : le modèle apprend l'échantillon par cœur, bruit compris.
- Capacité **trop petite** → **sous-apprentissage** : le modèle rate la structure.

Aucun des deux n'est visible sur l'erreur d'apprentissage seule. Le sur-apprentissage y est **invisible par construction** : $\hat{R}(f^*, D_n)$ décroît avec la capacité et ne remonte jamais.

54

Premier remède : pénaliser

$$f^* = \arg \min_{f \in \mathcal{F}} \hat{R}(f, D_n) + H(f)$$

où $H(f)$ croît avec la capacité de f .

On ne restreint plus $\mathcal{F}$: on rend la complexité **coûteuse**.

- L'élagage des arbres CART et
- la régression ridge sont deux exemples de pénalisation de la complexité.

55

Second remède : arrêter à temps

Applicable **si l'apprentissage est itératif**.

Au lieu de minimiser le risque empirique jusqu'au bout :

1. à chaque itération, estimer l'erreur de généralisation ;
2. **arrêter quand elle cesse de diminuer**.

C'est le principe de l'*early stopping*, omniprésent en apprentissage profond. Il exige des données distinctes de celles qui servent à ajuster le modèle afin d'estimer l'erreur de généralisation.

56

En résumé

1. Le risque empirique est sans biais **à f fixée**, plus après optimisation.
2. L'erreur d'apprentissage est **optimiste** : ce n'est pas un résultat.
3. L'erreur se décompose en **biais² + variance + bruit**.
4. Augmenter la capacité échange du biais contre de la variance.
5. Deux remèdes : **pénaliser** la complexité, ou **arrêter tôt**.

57

En pratique

58

Trois problèmes distincts

1. **Choisir la complexité** de $\mathcal{F}$ — sélection de modèle
2. **Ajuster les paramètres** à complexité fixée
3. **Estimer le risque** de la fonction retenue

Ce sont trois questions distinctes. Les traiter sur les mêmes données est l'erreur méthodologique la plus courante — et la plus coûteuse.

59

Les dangers du « en même temps »

- En théorie, les risques sont des **espérances**.
- En pratique, on manipule des **moyennes** sur un échantillon fini.
- Chaque fois qu'on utilise les mêmes données pour décider **et** pour évaluer, on se spécialise un peu plus sur elles.

Tout jeu de données ayant servi à prendre une décision est discrédité pour l'évaluation.

60

Le découpage

Trois rôles distincts — un seul jeu ne peut pas les tenir tous



Figure 12

Un jeu par rôle. C'est le minimum méthodologique.

61

Estimation du risque par ensemble de test

Partage

$$D_n^{tr} = \{z_1, \dots, z_{tr}\}, \quad D_n^{te} = D_n \setminus D_n^{tr}$$

Procédure

1. Sélectionner $f^*(D_n^{tr}) = \arg \min_{f \in \mathcal{F}} \hat{R}(f, D_n^{tr})$
2. Estimer $R(f)$ par

$$\hat{R}(f^*(D_n^{tr}), D_n^{te}) = \frac{1}{|D_n^{te}|} \sum_{z_i \in D_n^{te}} L(z_i, f^*(D_n^{tr}))$$

62

Le problème du découpage simple

Avec les 150 fleurs d'Iris et un test à 30 %, l'estimation repose sur **45 observations**.

Deux conséquences :

- l'estimation est **bruitée** — changez la graine, changez le résultat ;
- les données de test sont **perdues** pour l'apprentissage.

63

Validation croisée

Validation croisée à 5 plis : chaque observation sert une fois

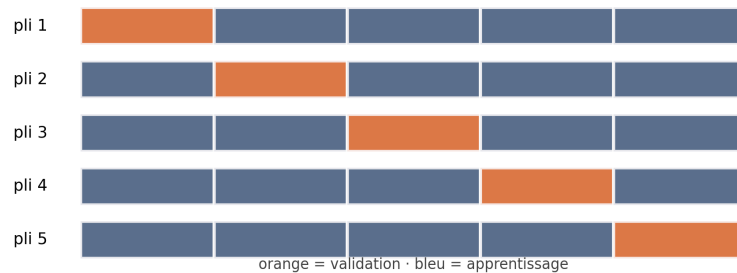


Figure 13

Les plis sont de taille approximativement égale, et **stratifiés** quand les classes sont déséquilibrées.

64

Validation croisée – la procédure

Partage de D_n en K plis $D_1, \dots, D_K$. On note $\bar{D}^k = D_n \setminus D_k$.

Pour chaque k :

- sélectionner $f^*(\bar{D}^k) = \arg \min_{f \in \mathcal{F}} \hat{R}(f, \bar{D}^k)$
- estimer $\hat{R}(f^*(\bar{D}^k), D_k) = \frac{1}{|D_k|} \sum_{z_i \in D_k} L(z_i, f^*(\bar{D}^k))$

Puis moyenner :

$$\hat{R}_{CV}(f) = \frac{1}{K} \sum_{k=1}^K \hat{R}(f^*(\bar{D}^k), D_k)$$

Chaque observation sert **une fois** à la validation et $K - 1$ fois à l'apprentissage.

65

Une hypothèse cachée

- Le découpage aléatoire suppose les observations **interchangeables**.
- Faux avec une dépendance temporelle : Deux relevés espacés de cinq minutes sont presque le même relevé.

Répartir au hasard des observations voisines place des quasi-jumelles de part et d'autre du découpage. Le modèle retrouve en test ce qu'il a mémorisé en apprentissage.

66

Deux découpages, trois chiffres

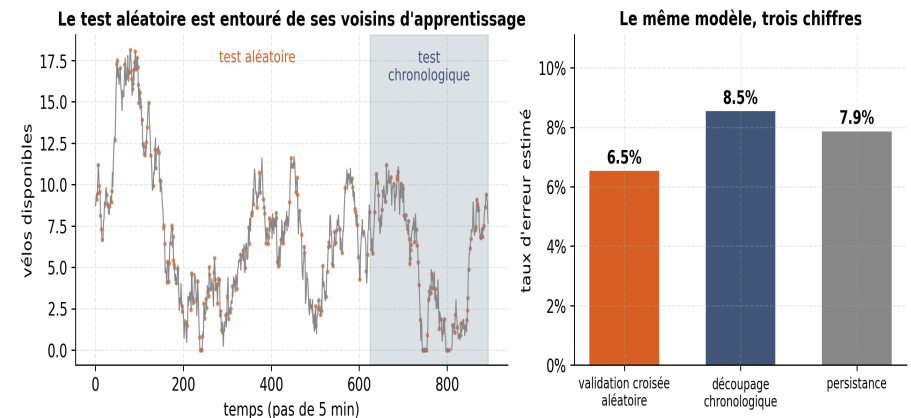


Figure 14

Parc de stations simulé. Cible : station en tension trente minutes plus tard.

67

Ce que la figure établit

La validation croisée aléatoire annonce **6,5 %**. Le découpage chronologique mesure **8,5 %**.

Deux points d'écart, sur le même modèle et les mêmes données. Le premier chiffre est faux.

Quand les observations se suivent dans le temps, le jeu de test est ce qui vient **après**. Jamais un tirage au hasard.

68

La référence de persistance

- Prédire que l'état futur sera l'état actuel. Aucun apprentissage.
- Elle atteint **7,9 %** — mieux que le modèle honnêtement évalué.

Le classifieur constant est la référence minimale. Sur des données temporelles, la persistance en est une seconde, souvent bien plus exigeante.

Un modèle qui ne la bat pas n'a rien appris d'utile.

69

Sélectionner et estimer, simultanément

Pour à la fois **trouver la bonne complexité** et **estimer le risque** de la fonction retenue, il faut **trois** ensembles disjoints :

- **apprentissage** — ajuster les paramètres à complexité fixée
- **validation** — choisir la complexité
- **test** — estimer le risque final, une seule fois

Si vous avez regardé le jeu de test plus d'une fois, ce n'est plus un jeu de test.

70

Petits jeux de données

Stratégies pour les pauvres :

- une **validation croisée** pour choisir la complexité ;
- une **double validation croisée** (imbriquée) pour choisir la complexité et estimer son risque.

Avec $n = 150$, la double validation croisée limite le gaspillage de données tout en séparant la sélection du modèle de l'estimation de son risque.

71

Un taux d'erreur ne suffit pas

Sur les scores simulés qui servent aux figures suivantes — 35 % de malades —, répondre toujours « sain » donne **64,8 %** de bonnes réponses.

Le chiffre est flatteur. Le classifieur ne détecte **aucun** malade.

Sous déséquilibre, l'exactitude globale récompense la classe majoritaire. Il faut compter les erreurs séparément.

72

La matrice de confusion

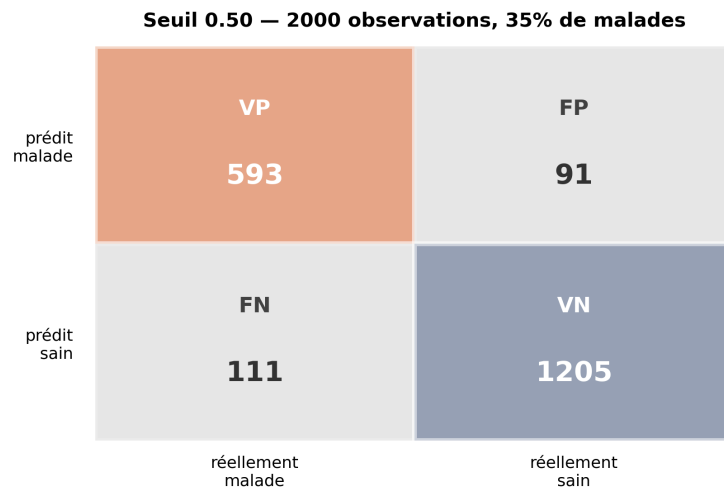


Figure 15

Quatre effectifs, obtenus en croisant la vérité et la décision.

73

Réalité versus prédiction

	Réellement malade	Réellement sain
Prédit malade	VP – vrai positif	FP – faux positif
Prédit sain	FN – faux négatif	VN – vrai négatif

Les deux diagonales suffisent à tout reconstruire :

$$n = VP + FP + FN + VN, \quad \text{exactitude} = \frac{VP + VN}{n}.$$

74

Deux points de vue

Conditionner par la **vérité** :

$$TPR = \frac{VP}{VP + FN}, \quad FPR = \frac{FP}{FP + VN}.$$

Le TPR s'appelle aussi **rappel**, ou sensibilité.

Conditionner par la **prédiction** :

$$\text{précision} = \frac{VP}{VP + FP}, \quad FDR = \frac{FP}{VP + FP} = 1 - \text{précision}$$

Le rappel répond : *quelle part des malades ai-je trouvée ?* La précision répond : *quelle part de mes alertes était fondée ?*

75

Deux questions différentes

Un dépistage qui déclare **tout le monde malade** atteint un rappel de 100 %.

Sa précision tombe à la prévalence – 35 % ici. Deux alertes sur trois sont fausses.

À l'inverse, n'alerter que sur le cas le plus certain donne une précision proche de 1 et un rappel proche de 0.

Chacune se maximise trivialement au détriment de l'autre.

76

Le score F1

Moyenne **harmonique** de la précision et du rappel :

$$F_1 = 2 \frac{\text{précision} \times \text{rappel}}{\text{précision} + \text{rappel}} = \frac{2 \text{ VP}}{2 \text{ VP} + \text{FP} + \text{FN}}.$$

La moyenne harmonique s'effondre dès qu'un des deux termes est faible. Une précision de 1 et un rappel de 0,02 donnent $F_1 \approx 0,04$, là où la moyenne arithmétique afficherait 0,51.

F_1 ignore les vrais négatifs. Sur un problème très déséquilibré, c'est un avantage, pas un oubli.

77

Le modèle ne change pas. Seule la valeur au-delà de laquelle on décide « malade » change.

78

Tout dépend du seuil

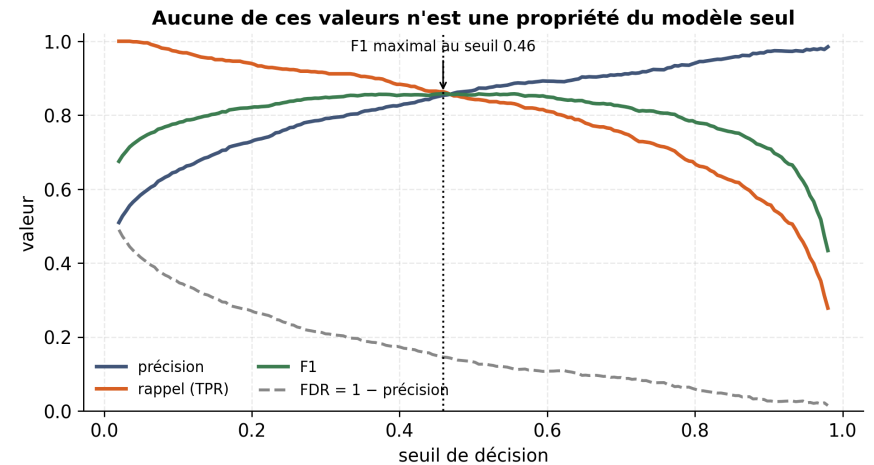


Figure 16

Interpréter les métriques

- Annoncer une précision sans annoncer le seuil ne veut rien dire.
- Le seuil 0,5 n'a rien de particulier : il correspond à la règle MAP, donc à l'hypothèse implicite que les deux erreurs ont le même coût.

Le seuil est un paramètre de décision, pas un paramètre du modèle. Il se choisit à partir des coûts.

79

La courbe ROC

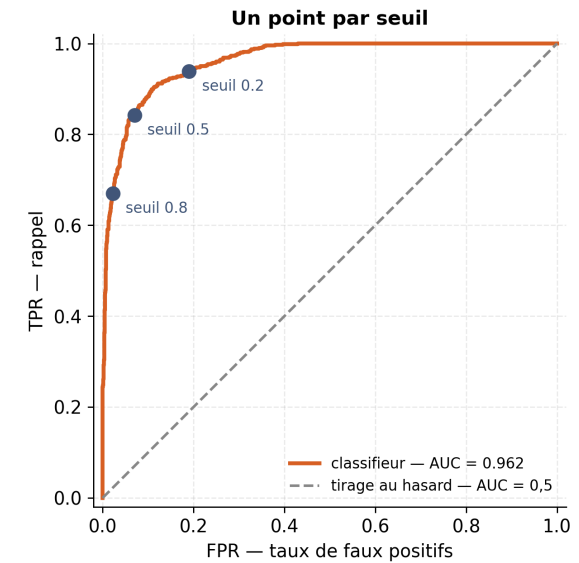


Figure 17

Lire une courbe ROC

- Le coin supérieur gauche est le classifieur parfait
- La diagonale est le tirage au hasard.

L'**AUC** est l'aire sous la courbe. Elle vaut la probabilité qu'un malade tiré au hasard reçoive un score supérieur à celui d'un sain.

Une seule valeur, indépendante du seuil et de la prévalence.

La courbe précision-rappel

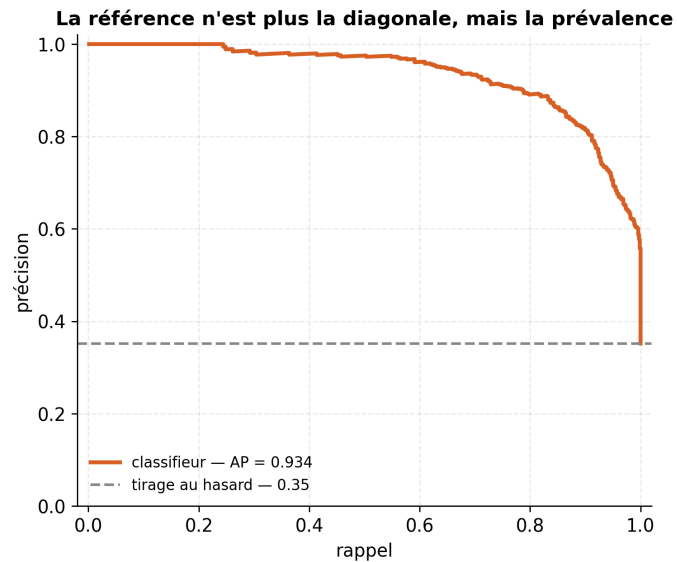


Figure 18

82

Même modèle, deux prévalences. L'AUC bouge à peine, la précision moyenne s'effondre.

83

Laquelle choisir

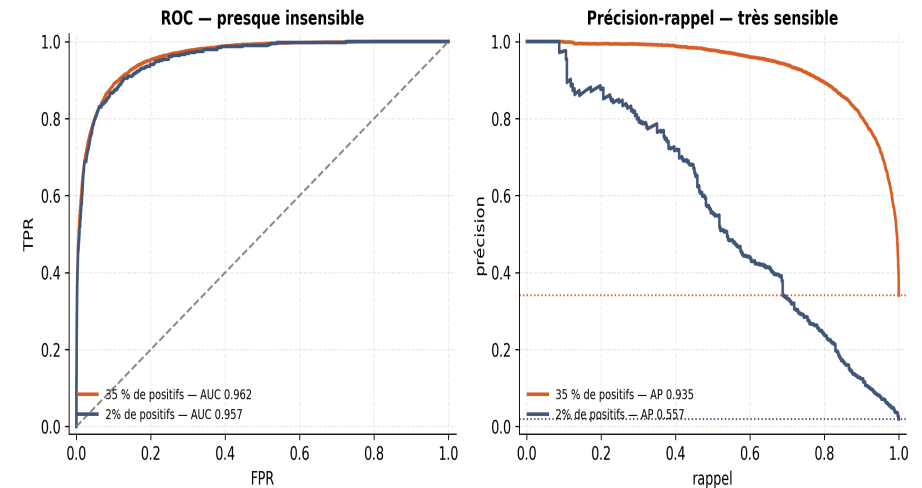


Figure 19

La raison

- La ROC conditionne par la vérité : **TPR** et **FPR** ont chacun leur propre dénominateur. Changer la proportion de positifs ne les déplace pas.
- La précision, elle, mélange les deux classes à son dénominateur. Elle chute dès que les négatifs deviennent nombreux

- **Classes équilibrées**, ou comparaison de modèles entre eux : la ROC suffit.
- **Positifs rares**, et coût réel des fausses alertes : la courbe précision-rappel est la seule honnête.

84

En pratique

```
1 from sklearn.metrics import (confusion_matrix, classification_report,
2                               roc_auc_score, average_precision_score)
3
4 confusion_matrix(y_test, y_pred)
5 classification_report(y_test, y_pred)    # précision, rappel, F1 par classe
6
7 proba = modele.predict_proba(X_test)[: , 1]
8 roc_auc_score(y_test, proba)            # AUC
9 average_precision_score(y_test, proba)   # aire sous précision-rappel
```

Les deux dernières prennent un **score**, pas une décision. Leur passer **y_pred** revient à n'évaluer qu'un seul seuil.

85

Le fléau de la dimension

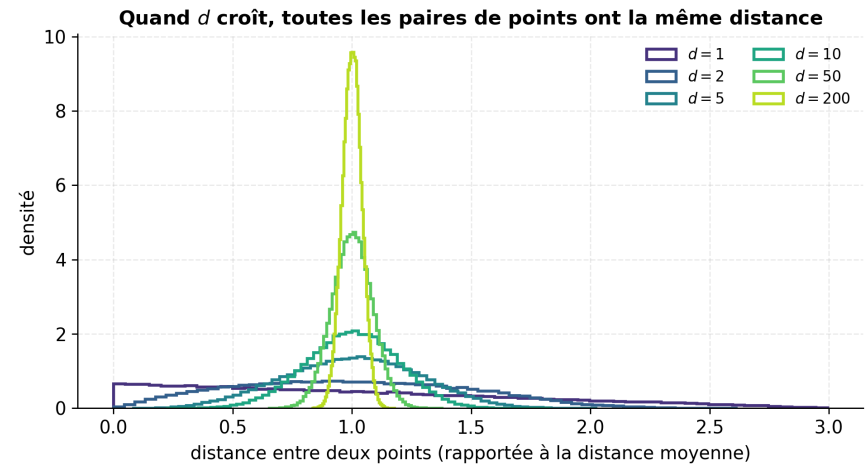


Figure 20

En grande dimension, toutes les paires de points ont sensiblement la même distance. La notion de « voisin » se vide de son sens.

86

Le fléau de la dimension

- Les méthodes fondées sur une **distance** (k -PPV, noyaux) perdent leur fondement.
- Le volume à couvrir croît exponentiellement : à densité d'échantillonnage constante, n devrait croître comme c^d .
- Ajouter une variable non informative **dégrade** le modèle.

87

En bref

1. Apprendre = minimiser un risque qu'on ne sait pas calculer.
2. L'erreur d'apprentissage est optimiste ; seule une donnée jamais vue mesure quelque chose.
3. Biais et variance s'échangent ; le compromis est forcé.
4. Trois rôles, trois jeux de données. Un jeu ayant servi à décider est discrédité.
5. Le coût est un choix, pas une donnée.