EE-311—Apprentissage et intelligence artificielle

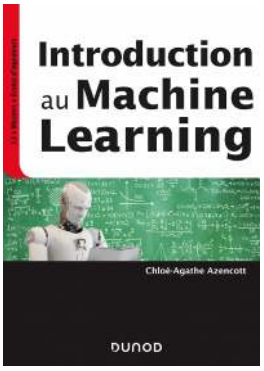
11. Sélection de modèle et évaluation

Michael Liebling

https://moodle.epfl.ch/course/view.php?id=16090

16 mai 1024 (compilé le 13 mai 2025)

Ouvrage de référence et source



Ces transparents sont basés en grande partie sur le texte de Chloé-Agathe Azencott "Introduction au Machine Learning", Dunod, 2019
ISBN 978-210-080153-4

L'auteure a mis le texte (sans les exercices) à disposition ici :
http://cazencott.info/dotclear/public/lectures/IntroML_Azencott.pdf

Avertissement : Bien que ces transparents partagent la notation mathématique, la structure de l'exposition (en partie), et certains exemples avec le livre, ils ne constituent qu'un complément et non un remplacement ou une source unique pour la couverture des matières du cours. À ce titre, ces transparents ne se substituent pas au texte.

Contenu

- Sur- et sous-apprentissage, généralisation (rappel)
- Dilemne biais-variance
- Évaluation des méthodes d'apprentissage :
  - concevoir un cadre expérimental dans lequel sélectionner un modèle d'apprentissage supervisé
  - choisir un ou des critères d'évaluation d'un modèle d'apprentissage supervisé
  - estimer la performance en généralisation d'un modèle d'apprentissage supervisé

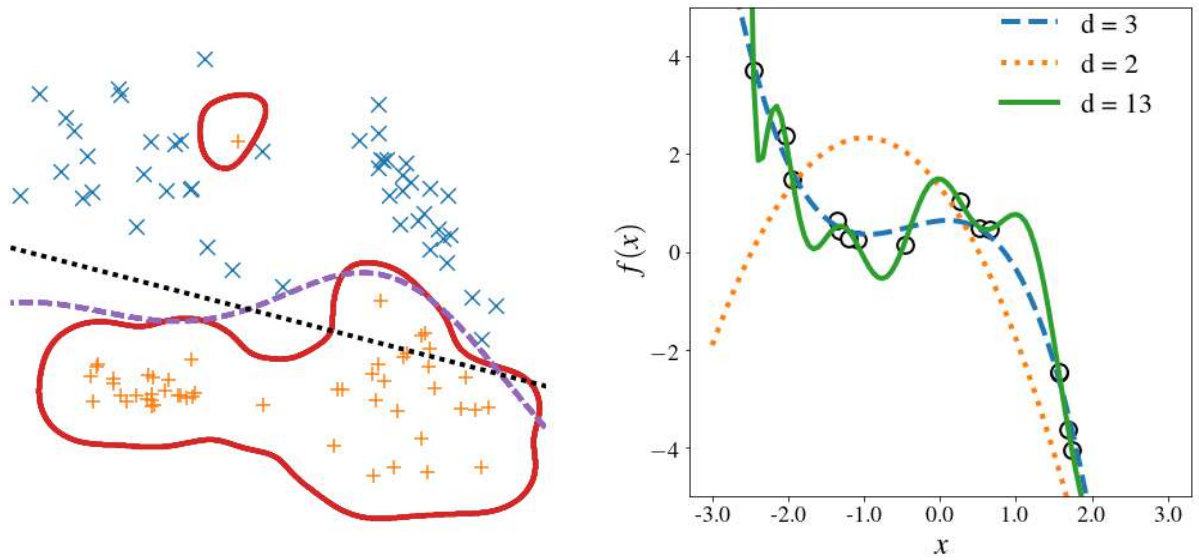
Généralisation et sur-apprentissage (rappel du Cours 2)

Définition 2.21 (Généralisation) On appelle généralisation la capacité d'un modèle à faire des prédictions correctes sur de nouvelles données, qui n'ont pas été utilisées pour le construire.

Définition 2.22 (Sur-apprentissage) On dit d'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. En anglais, on parle d'overfitting.

Définition 2.23 (Sous-apprentissage) On dit d'un modèle qui est trop simple pour avoir de bonnes performances même sur les données utilisées pour le construire qu'il sous-apprend. En anglais, on parle d'underfitting.

Illustration : Sous-apprentissage et sur-apprentissage (rappel du Cours 2)



(A) Pour séparer les observations négatives (x) des observations positives (+), la droite pointillée sous-apprend. La frontière de séparation en trait plein ne fait aucune erreur sur les données mais est susceptible de sur-apprendre. La frontière de séparation en trait discontinu est un bon compromis.

(B) 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, sur-apprend.

FIGURE 2.6 – Sous-apprentissage et sur-apprentissage

Azencott

Compromis biais-variance

Pour mieux comprendre le risque d'un modèle $f : \mathcal{X} \to \mathcal{Y}$, nous pouvons le comparer à l'erreur minimale $\mathscr{R}^*$ qui peut être atteinte par n'importe quelle fonction mesurable de $\mathcal{X}$ dans $\mathcal{Y}$: c'est ce qu'on appelle l'*excès d'erreur*, et que l'on peut décomposer de la façon suivante :

$$\mathscr{R}(f) - \mathscr{R}^* = \underbrace{\left[\mathscr{R}(f) - \min_{h \in \mathcal{F}} \mathscr{R}(h)\right]}_{\text{Erreur d'estimation}} + \underbrace{\left[\min_{h \in \mathcal{F}} \mathscr{R}(h) - \mathscr{R}^*\right]}_{\text{Erreur d'approximation}}$$

Erreur d'estimation :
distance entre modèle f et le modèle optimal sur $\mathcal{F}$
"variance"

Erreur d'approximation :
la qualité du modèle (dans $\mathcal{F}$) optimal
qualité du choix de l'espace des hypothèses
"biais"

Example : approximation d'une fonction sinusoidale par des B-splines

On veut approximer une fonction sinusoïdale

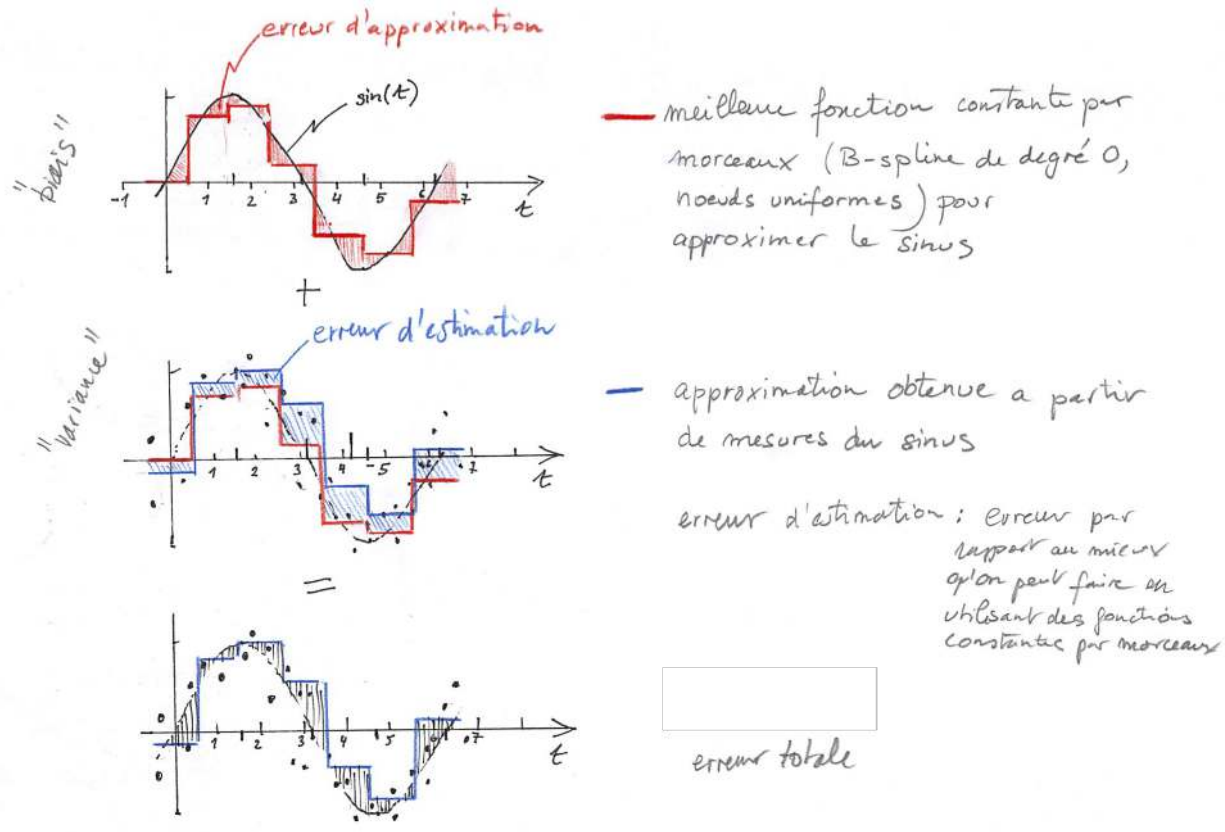
$$x(t) = \sin(\alpha t)$$

par une fonction constante par morceaux, discontinues entre deux entiers (B-splines centrés de degré 0)

$$\tilde{x}(t) = \sum_{k \in \mathbb{Z}} c_k 1_{[-\frac{1}{2}, \frac{1}{2}]}(t - k).$$

On cherchera à estimer les coefficients c_k.

Illustration compromis biais-variance



Remarques dilemne biais-variance

- Il est clair qu'en fixant ce modèle simpliste (fonction escalier) on ne pourra jamais modéliser le sinus exactement (on fait une approximation *biaisée* par ce choix de modèle).
- L'*erreur d'approximation* caractérise la meilleure approximation possible (biaisée)
- En pratique, comme on fait l'estimation des paramètres sur des mesures, on n'obtiendra en général même pas la qualité de cette approximation : on fait, en plus, une *erreur d'estimation*.
- erreur totale = erreur d'approximation + erreur d'estimation.
- En choisissant un modèle plus "précis" (par example avec des noeuds plus serrés) on réduira bien l'erreur d'approximation mais pas forcément l'erreur d'estimation (car on a plus de paramètres à estimer mais toujours les mêmes données).
- Le modèle avec un plus grand biais pourra dans certains cas s'avérer préférable pour obtenir une erreur totale moindre !

Estimation de densité : décomposition biais-variance

Lorsqu'on désire estimer le paramètre θ d'une densité, l'erreur quadratique moyenne d'un estimateur $\hat{\theta}$ est donnée par

$$
\begin{aligned}
\mathrm{MSE}\left(\hat{\theta}\right) &= \mathbb{E}\left[\left(\hat{\theta}-\theta\right)^2\right] \\
&= \mathbb{E}\Big[\big(\hat{\theta}\underbrace{-\mathbb{E}\left[\hat{\theta}\right]+\mathbb{E}\left[\hat{\theta}\right]}_{+0}-\theta\big)^2\Big] \\
&= \mathbb{E}\left[\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)^2\right] + \mathbb{E}\left[\underbrace{\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)}_{\substack{\text{déterministe :}\\ \text{fixe}}}{}^2\right] \\
&\qquad + \mathbb{E}\left[2\left(\hat{\theta}-\mathbb{E}\left[\hat{\theta}\right]\right)\underbrace{\left(\mathbb{E}\left[\hat{\theta}\right]-\theta\right)}_{\substack{\text{déterministe :}\\ \text{quantité}\\ \text{fixe et}\\ \text{indépendante}}}\right] \\
&= \cdots \quad \text{(page suivante)}
\end{aligned}
$$

Décomposition biais-variance d'un estimateur de paramètre de densité (suite)

$$\mathrm{MSE}\left(\hat{\theta}\right) = \cdots = \mathrm{Var}(\hat{\theta}) + \left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2 + \mathbb{E}\left[2\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\right] \underbrace{\mathbb{E}\left[\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)\right]}_{\mathbb{E}\left[\mathbb{E}\left[\hat{\theta}\right]\right] - \mathbb{E}[\theta] = 0}$$

$$= \underbrace{\mathrm{Var}\left(\hat{\theta}\right)}_{\text{variance}} + \underbrace{\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2}_{\text{carré du biais}}$$

$\Rightarrow$ un estimateur biaisé peut, si sa variance est plus faible, avoir une erreur quadratique moyenne plus faible qu'un estimateur non biaisé.

C'est une nouvelle manifestation de la notion de compromis biais-variance !

Rappel : Régression linéaire (sans offset)

Modèle linéaire (par exemple pour une fonction de décision) :

$$f : \vec{x} \mapsto \sum_{j=1}^{p} \beta_j x_j = \begin{pmatrix} x_1 & \dots & x_p \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \vec{x}\vec{\beta}$$

avec $\vec{x} \in \mathbb{R}^{1\times p}$ et $\vec{\beta} \in \mathbb{R}^{p\times 1}$ (Nombre de variables : p).

Régression linéaire par minimisation des moindres carrés : on cherche le modèle de la forme $f : \vec{x} \mapsto \sum_{j=1}^{p} \beta_j x_j$ dont les coefficients sont obtenus par :

$$\vec{\beta}^*_{\mathrm{LS}} = \arg\min_{\vec{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \left(y^i - \sum_{j=1}^{p} \beta_j x_j^i \right)^2$$

Solution : Si X est de rang égal à son nombre de colonnes p, on a :

$$\vec{\beta}^*_{\mathrm{LS}} = \left(X^\top X\right)^{-1} X^\top \vec{y}$$

$\vec{\beta}^*_{LS}$ est un estimateur non-biaisé de β

Régression Ridge

Définition (régression ridge) : On appelle *régression ridge* le modèle

$$f : \vec{x} \mapsto \sum_{j=1}^{p} \beta_j x_j = \begin{pmatrix} x_1 & \dots & x_p \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} = \vec{x}\vec{\beta}$$

dont les coefficients sont : $f : \vec{x} \mapsto \sum_{j=1}^{p} \beta_j x_j$ dont les coefficients sont obtenus par :

$$\vec{\beta}^*_{\text{ridge}} = \arg\min_{\vec{\beta} \in \mathbb{R}^p} \left\| \vec{y} - X\vec{\beta} \right\|_2^2 + \lambda \left\| \vec{\beta} \right\|_2^2$$

Solution : (pas de condition sur le rang de X) on a :

$$\vec{\beta}^*_{\text{ridge}} = \left(\lambda I_p + X^\top X\right)^{-1} X^\top \vec{y}$$

$\vec{\beta}^*_{ridge}$ *est un estimateur biaisé de* β

Un estimateur biaisé peut avoir des avantages comparé à un estimateur non-biaisé

Avantages de l'estimateur ridge (estimateur biaisé) :

- régularisation du problème : si la matrice $(X^\top X)$ est mal conditionnée, calculer son inverse (estimateur moindre carrés) sera instable, contrairement à l'expression $\left(\lambda I_p + X^\top X\right)^{-1}$ de l'estimateur ridge
- bien que l'estimateur ridge soit biaisé, comme l'erreur est la somme du biais au carré et de la variance, si la variance est faible (grâce à une inversion plus stable), l'erreur totale pourrait ainsi être plus faible que pour l'estimateur non-biaisé
- lorsqu'on a n'a pas suffisemment de mesures $n < p$ la régression ridge permet néanmoins d'obtenir une estimation des paramètres

Évaluation $\neq$ sélection

Comment mettre en place un cadre expérimental qui permette d'évaluer un modèle en évitant le biais du sur-apprentissage ?

Distinguer :

- évaluation d'un modèle, qui consiste à déterminer sa performance sur l'espace des données dans sa totalité
- sélection du modèle, qui consiste à choisir le meilleur modèle parmi plusieurs.

Estimation empirique de l'erreur de généralisation (rappel cours 5)

L'erreur empirique mesurée sur les observations qui ont permis de construire le modèle est un mauvais estimateur de l'erreur du modèle sur l'ensemble des données possibles, ou erreur de généralisation : si le modèle sur-apprend, cette erreur empirique peut être proche de zéro voire nulle, tandis que l'erreur de généralisation peut être arbitrairement grande.

Pour évaluer la qualité d'un modèle appris, on sépare communément les données en trois jeux de données (pourcentages indicatifs, règle générale) :

1. jeu d'entraînement (60-70% des données)
2. jeu de validation (15-20% des données), e.g. si plusieurs modèles sont considérés ou si le modèle à entraîner a des paramètres
3. jeu de test (15-20% des données)

Jeu d'entraînement, Jeu de test (rappel cours 5)

Pour évaluer un modèle, il est indispensable d'utiliser des données étiquetées qui n'ont **pas** *servi à le construire.*

Définition 3.1 (Jeu d'entraînement, Jeu de test) Étant donné un jeu de données $\mathscr{D} = \left\{\left(\vec{x}^i, y^i\right)\right\}_{i=1,\ldots,n}$ partitionné en deux jeux $\mathscr{D}_{\text{tr}}$ et $\mathscr{D}_{\text{te}}$, on appelle jeu d'entraînement (training set en anglais) l'ensemble $\mathscr{D}_{\text{tr}}$ utilisé pour entraîner un modèle prédictif, et jeu de test (test set en anglais) l'ensemble $\mathscr{D}_{\text{te}}$ utilisé pour son évaluation. **La perte calculée sur ce jeu de test est un estimateur de l'erreur de généralisation.**

Attention : manquer à séparer les jeux d'entraînement et de test (e.g. en présentant comme la performance d'un modèle son erreur sur le jeu d'entraînement) est probablement le pêché capital du machine learning !

Jeu de validation (rappel cours 5)

Considérons la situation où nous devons choisir entre K modèles : nous pouvons entraîner chacun des modèles sur le jeu de données d'entraînement, obtenant ainsi K fonctions de décision $f_1, f_2, \ldots, f_K$.

Comment choisir le meilleur modèle ? Si on calcule l'erreur de chacun de ces modèles sur le jeu de test pour choisir le meilleur, nous ne pourrons plus utiliser le jeu de test pour évaluer l'erreur de généralisation du modèle choisi.

Plutôt, nous définissons un jeu de validation $\mathscr{D}_{\text{val}}$, sur lequel on peut choisir le modèle qui a la plus petite erreur :

$$\hat{f} = \underset{k=1,\ldots,K}{\operatorname{argmin}} \frac{1}{|\mathscr{D}_{\text{val}}|} \sum_{\vec{x},y \in \mathscr{D}_{\text{val}}} L\left(y, f_k(\vec{x})\right)$$

Importance de distinguer la sélection d'un modèle de son évaluation : les faire sur les mêmes données peut nous conduire à sous-estimer l'erreur de généralisation et le sur-apprentissage du modèle choisi.

Solution de découpage des jeux de données (rappel cours 5)

Entraînement d'un seul modèle sans paramètre

1. jeu d'entraînement $\mathscr{D}_{\text{tr}}$ sur lequel on entraîne l'algorithme d'apprentissage
2. jeu de test $\mathscr{D}_{\text{te}}$ sur lequel on évalue l'erreur de généralisation du modèle

Entraînement d'un modèle avec paramètres ou lorsque le modèle doit être choisi parmi plusieurs

1. jeu d'entraînement $\mathscr{D}_{\text{tr}}$ sur lequel on entraîne K algorithmes d'apprentissage
2. jeu de validation $\mathscr{D}_{\text{val}}$ sur lequel on évalue les K modèles pour sélectionner le modèle définitif
3. jeu de test $\mathscr{D}_{\text{te}}$ sur lequel on évalue l'erreur de généralisation du modèle choisi.

Validation croisée (rappel cours 5)

Définition 3.2 (Validation croisée) Étant donné un jeu $\mathscr{D}$ de n observations, et un nombre K, on appelle validation croisée la procédure qui consiste à

1. partitionner $\mathscr{D}$ en K parties de tailles sensiblement similaires, $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_K$
2. pour chaque valeur de $k = 1, \ldots, K$,
   - entraîner un modèle sur $\bigcup_{\ell \neq k} \mathscr{D}_\ell$
   - évaluer ce modèle sur $\mathscr{D}_k$.

Chaque partition de $\mathscr{D}$ en deux ensembles $\mathscr{D}_k$ et $\bigcup_{\ell \neq k} \mathscr{D}_\ell$ est appelée un fold de la validation croisée.

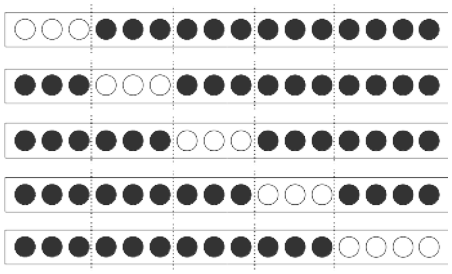


FIGURE 3.1 – Une validation croisée en 5 *folds* : Chaque observation appartient à un des 5 jeux de validation (en blanc) et aux 4 autres jeux d'entraînement (en noir).

Azencott

Stratification (nouveau !)

Définition 3.3 (Validation croisée stratifiée) Une validation croisée est dite *stratifiée* si la moyenne des étiquettes des observations est sensiblement la même dans chacun des K sous-ensembles $\mathscr{D}_k$:

$$\frac{1}{|\mathscr{D}_1|}\sum_{i\in\mathscr{D}_1} y^i \approx \frac{1}{|\mathscr{D}_2|}\sum_{i\in\mathscr{D}_2} y^i \approx \cdots \approx \frac{1}{|\mathscr{D}_K|}\sum_{i\in\mathscr{D}_K} y^i \approx \frac{1}{|\mathscr{D}|}\sum_{i\in\mathscr{D}} y^i$$

Dans le cas d'un problème de classification, cela signifie que la proportion d'exemples de chaque classe est la même dans chacun des $\mathscr{D}_k$. Cette proportion est donc aussi la même que dans le jeu de données $\mathscr{D}$ complet.

L'intérêt de cette procédure est de faire en sorte que la distribution des observations au sein de chaque $\mathscr{D}_k$ soit la même qu'au sein du jeu de données $\mathscr{D}$.

Stratification : justification

Exemple : si par malchance un des folds ne contient que des exemples positifs dans son jeu d'entraînement et que des exemples négatifs dans son jeu de test, il est vraisemblable que, sur ce fold, tout le modèle apprenne à prédire que tout est positif et ait une très mauvaise performance.

Validation croisée leave-one-out

Définition 3.4 (Validation croisée leave-one-out) Une validation croisée dont le nombre de folds est égal au nombre d'observations dans le jeu d'entraînement, et dont chaque fold est donc composé d'un jeu d'entraînement de taille $n-1$ et d'un jeu de test de taille 1, est appelée *leave-one-out* : on met de côté, pour chaque fold, un unique exemple.

Intuition un algorithme d'apprentissage apprendra d'autant mieux qu'il y a d'avantage de données disponibles pour l'entraînement : plus on connaît d'étiquettes pour des observations de l'espace $\mathcal{X}$, plus on peut contraindre le modèle à les respecter. Or pour un jeu de données de taille n, un jeu de test d'une validation croisée à K folds contient $\frac{(K-1)n}{K}$ points : les modèles entraînés apprendront d'autant mieux sur chacun des folds qu'ils sont grands, ce qui nous pousse à considérer le cas où $K = n$.

Inconvénients du leave-one-out

Le leave-one-out a deux inconvénients :

- requiert un grand temps de calcul (on entraîne n modèles, chacun sur $n-1$ observations au lieu de (dans le cas $K = 10$) 10 modèles, chacun sur 90% des observations)
- les jeux d'entraînements ainsi formés sont très similaires entre eux et les modèles entraînés le seront aussi, et peu différents d'un modèle entraîné sur l'intégralité du jeu de données.
- les jeux de test seront disjoints, et les performances pourront avoir une grande variabilité (interprétation plus compliquée)

Bootstrap

But : rééchantillonner les données afin d'estimer l'erreur de généralisation

Définition 3.5 (Bootstrap) Étant donné un jeu $\mathscr{D}$ de n observations et un nombre B, on appelle *bootstrap* la procédure qui consiste à créer B échantillons $\mathscr{D}_1, \mathscr{D}_2, \ldots, \mathscr{D}_B$ de $\mathscr{D}$, obtenus chacun en tirant n exemples de $\mathscr{D}$ *avec remplacement*. Ainsi chaque exemple peut apparaître plusieurs fois, ou pas du tout, dans $\mathscr{D}_b$.
Remarque : le bootstrap est une procédure couramment utilisée en statistiques pour estimer un paramètre en fonction de son estimation sur les B échantillons.
Procédure :

1. entraînement du modèle à évaluer sur chaque échantillon $\mathscr{D}_b$
2. évaluation de sa performance sur l'intégralité de $\mathscr{D}$ (mieux : sur $\mathscr{D} \setminus \mathscr{D}_b$ pour éviter les biais)

Bootstrap discussion

Problème : si l'évaluation est faite sur $\mathscr{D}$, cette estimation serait biaisée par la présence d'une partie des exemples de $\mathscr{D}$ dans $\mathscr{D}_b$. Il faut donc se limiter aux exemples de $\mathscr{D} \setminus \mathscr{D}_b \Rightarrow$ procédure trop complexe en pratique

Remarque : La probabilité que $(\vec{x}^i, y^i)$ apparaisse dans $\mathscr{D}_b$ peut être calculée comme le complémentaire à 1 de la probabilité que $(\vec{x}^i, y^i)$ ne soit tiré aucune des n fois. La probabilité que $(\vec{x}^i, y^i)$ soit tiré une fois vaut $\frac{1}{n}$. Ainsi

$$\mathbb{P}\left[(\vec{x}^i, y^i) \in \mathscr{D}_b\right] = 1 - \left(1 - \frac{1}{n}\right)^n.$$

Quand n est grand, cette probabilité vaut donc environ $1 - \mathrm{e}^{-1} \approx 0.632$, car la limite en $+\infty$ de $\left(1 + \frac{x}{n}\right)^n$ vaut e^x.

Ainsi, $\mathscr{D}_b$ contient environ deux tiers des observations de $\mathscr{D}$.

Critères de performance

L'évaluation de la performance prédictive d'un modèle d'apprentissage supervisé peut se faire de nombreuses manières.

- Matrice de confusion, précision, rappel, F-mesure et spécificité
- Courbe ROC, courbe précision-rappel
- Erreurs de régression

Matrice de confusion

Définition 3.6 (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 classe, et dont l'entrée M_{ck} est le nombre d'exemples de la classe c pour laquelle l'étiquette k a été prédite.

Exemple (classification binaire)

		Classe réelle 0	Classe réelle 1
Classe prédite	0	vrais négatifs (TN)	faux négatifs (FN)
	1	faux positifs (FP)	vrais positifs (TP)

vrais positifs (true positives) : exemples $(+)$ correctement classifiés
faux positifs (false positives) : exemples $(-)$ classifiés comme $(+)$ par le modèle
vrais négatifs (true negatives) : exemples $(-)$ correctement classifiés
faux négatifs (false negatives) : exemples $(+)$ classifiés comme $(-)$ par le modèle

Exemple faux positifs/faux négatifs : examens de dépistage

Prédiction à partir d'une radiographie qu'une tumeur soit maligne (+) ou bénigne (-)

• Une prédiction positive (tumeur maligne) entraîne un examen approfondi (par exemple, une biopsie).

• Une prédiction négative (tumeur bénigne) n'entraîne pas d'examen supplémentaire.

Dépistage radiographique : peu invasif, fiabilité moindre

Examen approffondi par biopsie et analyse du tissu : invasif, plus fiable

→ Vrai positif : tumeur maligne, examen sera confirmé par examen approfondi (OK)

→ Vrai négatif : tumeur bénigne, pas d'examen approfondi ne sera effectué (OK)

→ Faux positif : tumeur bénigne, sera identifiée comme bénigne par examen approfondi (OK, mais ce dernier est coûteux et stressant)

→ Faux négatif : tumeur maligne, pas d'examen approfondi ne sera effectué et la tumeur maligne restera non-diagnostiquée ☹

Terminologies équivalentes

Faux positif (FP) : fausse alarme, erreur de type I

Faux négatifs (FN) : erreur de type II

Rappel

Définition 3.7 (Rappel) On appelle *rappel (recall)* ou *sensibilité (sensitivity)*, le taux de vrais positifs (*true positive rate, TPR*), i.e. la proportion d'exemples positifs correctement identifiés comme tels (par rapport au nombre d'exemples positifs présents) :

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

Rappelle-toi d'acheter...

Liste de courses d'articles ménagers : pain, lait, savon

Articles disponibles en magasin : pain, lait, pommes, chocolat, savon, papier toilette

Liste oubliée à la maison ! On passe en revue les rayons du magasin et on classifie chaque article suivant qu'on pense :

- qu'il faut l'acheter (présent sur la liste : positif), ou
- qu'il ne faut pas l'acheter (absent de la liste : négatif)

En rentrant à la maison, on peut évaluer :

Articles achetés et sur la liste : vrai positifs (TP)
Articles achetés et pas sur la liste : faux positifs (FP)
Articles non-achetés mais sur la liste : faux négatifs (FN)
Articles non-achetés et pas sur la liste : vrai négatifs (TN)

Illustration Rappel (Recall)

	🥖	🥛	🍎	🍫	🧼	🧻
🛒	+	+		+		
🚫			-		-	-
	TP	TP	TN	FP	FN	TN

Truc mnémotechnique :
Le rappel représente la proportion d'éléments de la liste de commissions qu'on s'est **rappel**é d'acheter.

Calcul du rappel :

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{2}{3}$$

Q : Comment peut-on obtenir un rappel parfait (=1) ?
Indice : la solution est coûteuse. . .

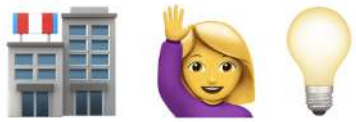
Rappel parfait : 1. se rappeler correctement de tout

	🥖	🥛	🍎	🍫	🧼	🧻
🛒	+	+			+	
🚫			-	-		-
	TP	TP	TN	TN	TP	TN

Calcul du rappel :

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3}{3} = 1$$

Rappel parfait : 2. tout acheter



	+	+	+	+	+	+
	TP	TP	FP	FP	TP	FP

Calcul du rappel :

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3}{3} = 1$$

Autre exemple (avec rappel = sensibilité = 1) : test de dépistage qui indique toujours un résultat positif (donc assuré de ne rater aucun cas de vraie malade)

Précision (precision)

Définiton 3.8 (Précision) on appelle *précision* la proportion de prédictions correctes parmi les prédictions positives :

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

Rappel et précision parfaits (se rappeler de tout correctem.)

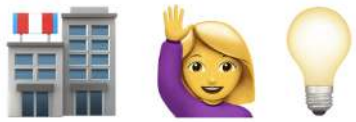
	🥖	🥛	🍎	🍫	🧼	🧻
🛒	+	+			+	
🚫			-	-		-
	TP	TP	TN	TN	TP	TN

Calcul du rappel et de la précision :

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3}{3} = 1 \quad ☺$$

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{3}{3} = 1 \quad ☺$$

Tout acheter : rappel parfait mais précision médiocre



	🥖	🥛	🍎	🍫	🧼	🧻
🛒	+	+	+	+	+	+
🚫						
	TP	TP	FP	FP	TP	FP

Calcul du rappel et de la précision :

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{3}{3} = 1 \quad ☺$$

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{3}{6} = 1/2 \quad ☹$$

Une astuce pour obtenir une bonne précision ?

Faire peu de prédictions positives (seulement acheter les choses dont on est absolument sûr qu'elles sont sur la liste)

		+				
	-		-	-	-	-
	FN	TP	TN	TN	FN	TN

$$\text{Rappel} = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{1}{1+2} = 1/3 \quad ☹$$

$$\text{Précision} = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{1}{1} = 1 \quad ☺$$

Accuracy and error rate

Pour obtenir une mesure de performance qui tient compte de toutes les prédictions (positives et négatives) on peut considérer :

Définiton 3.8b (Accuracy) on appelle *accuracy* la proportion de prédictions correctes parmi *toutes* les prédictions (positives *et* négatives) :

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

Définiton 3.8c (Taux d'erreur) on appelle *taux d'erreur* la proportion de prédictions incorrectes parmi *toutes* les prédictions (positives *et* négatives) :

$$\text{Taux d'erreur} = 1 - \text{Accuracy} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

F-mesure

Définition 3.9 (F-mesure) On appelle *F-mesure (F-score ou F1-score)* la moyenne harmonique de la précision et du rappel :

$$F = 2\frac{\text{Précision} \times \text{Rappel}}{\text{Précision}+\text{Rappel}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

Interprétation

- La F-mesure la plus haute est de 1.0, indiquant une parfaite précision et rappel.
- La F-mesure la plus basse est de 0, indiquant soit une précision ou un rappel zéro.

D'accord de payer plus pour oublier le moins de choses

Acheter trop en n'omettant que les choses dont on est sûr qu'elles ne sont pas sur la liste

	🥖	🥛	🍎	🍫	💵	🧻
🛒	+	+	+		+	
🚫				-		-
	TP	TP	FP	TN	TP	TN

Bonne spécificité :

$$\begin{aligned}\text{Spécificité} &= \frac{\text{TN}}{\text{FP} + \text{TN}} \\ &= \frac{2}{1+2} = 2/3\end{aligned}$$

Définition 3.10 (Spécificité) On appelle *spécificité* le taux de vrais négatifs, autrement dit, la proportion d'exemples négatifs correctement identifiés comme tels :

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

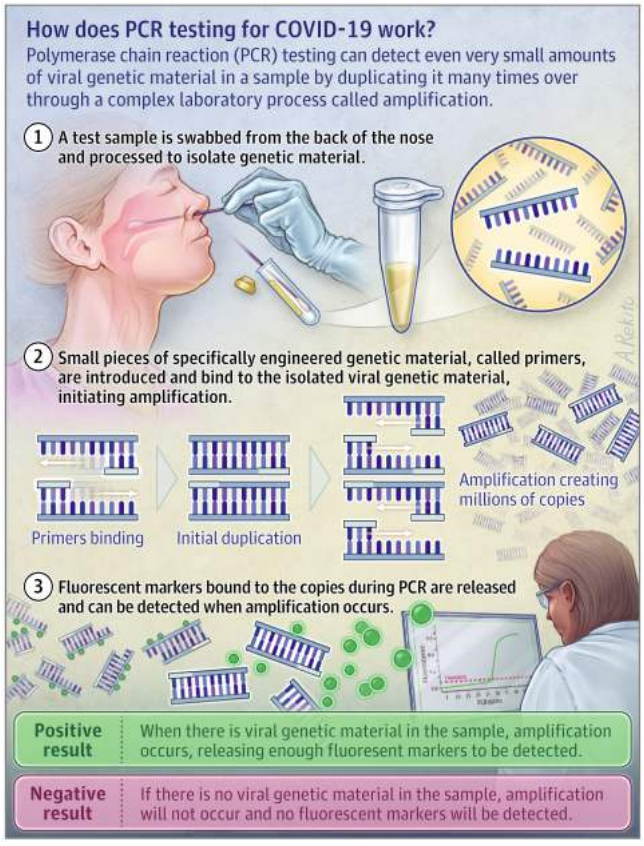
Exemple : test de dépistage d'une maladie

Pour identifier des personnes atteintes d'une maladie on a souvent accès à de multiple méthodes (questionnaire web, détection de présence d'anticorps, séquençage d'ADN, etc.) qui diffèrent par leur prix, rapidité, invasivité, précision et specificité.

Lors d'un dépistage, on cherche, en particulier, à :

- donner accès à un test à un grand nombre de personnes
  $\Rightarrow$ test peu coûteux, routinier, peu invasif, rapide
- s'assurer qu'on n'identifie (presque) aucune personne comme non-atteinte si elle est réellement atteinte de la maladie
  $\Rightarrow$ pas/peu de FN donc haut rappel
- identifier toutes les personnes potentiellement atteintes (+)
  $\Rightarrow$ précision pas critique, taux élevé de FP peut être acceptable
- mais ne pas identifier trop de personnes comme potentiellement atteintes (+) car les tests approfondis et précis sont souvent longs, coûteux, invasifs, anxiogènes (effets à long et court terme) $\rightarrow$ précision la plus haute possible sous contrainte de prix, temps, simplicité, invasivité

Dépistage coronavirus (analyse via réaction en chaîne par polymérase avec transcription inverse (RT-PCR))



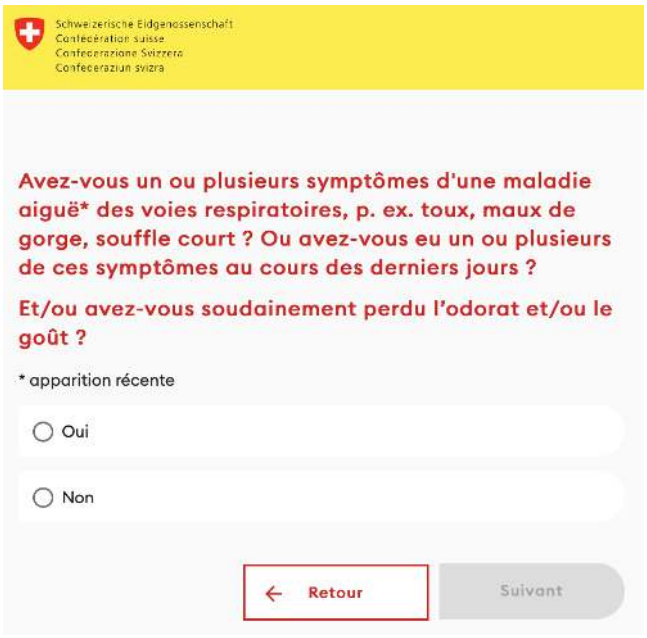
https ://cdn.jamanetwork.com/

Caractéristiques :

- peu-invasif (frottis nasal)
- requiert équipemment de laboratoire
- cher
- lent
- précis
- spécifique

Applied Biosystems

Coronavirus check (questionnaire web)



https://check.ofsp-coronavirus.ch/screening

Caractéristiques :

- non-invasif
- largement accessible
- peu coûteux
- rapide
- peu précis (symptômes subjectifs, similaires à d'autres maladies)
- test manquera personnes atteintes de Covid-19 ou porteuses du coronavirus mais asymptomatiques

Test clinique de dépistage du cancer du col de l'utérus

Frottis de dépistage : simple et relativement peu invasif
Prélèvement tissu et analyse histologique : invasif, cher, mais très fiable (vérité terrain)

Matrice de confusion

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

Rappel : 95% Spécificité : 94.5% Précision : 47.5% TN : 3590

→ Piètre précision : mauvais outil diagnostique (Les cas positifs requièrent un examen supplémentaire pour obtenir un diagnostic fiable.
→ Malgré la précision médiocre, bon test de dépistage : Probabilité de ne pas avoir le cancer si frottis est négatif est de $3590/3600 \approx 99.7\%$.

Résumé matrice de confusion

		Classe réelle		
		0 (N)	1 (P)	
Classe prédite	0 (N)	TN	FN	#(N pred)= TN + FN
	1 (P)	FP	TP	#(P pred)= FP + TP
	$\sum_{/\text{ diag}}$= #(F pred) = FP + FN	#(N) = TN + FP	#(P) = FN + TP	$\sum_{\backslash\text{diag}}$= #(T pred) = TN + TP

Matrice conf. : normalisation dans chaque classe réelle

		Classe réelle		
		0 (N)	1 (P)	
Classe prédite	0 (N)	Specificity : $\frac{TN}{TN+FP}$	1-Sensibility : $\frac{FN}{FN+TP}$	
	1 (P)	1-Specificity anti-specific. $\frac{FP}{TN+FP}$	Sensibility =Recall =rappel : $\frac{TP}{FN+TP}$	
		1	1	

Note : taille de la population totale et taille de chaque classe réelle ne sont plus lisibles avec cette normalisation.

Matrice conf. : normalisation dans chaque classe prédite

		Classe réelle		
		0 (N)	1 (P)	
Classe prédite	0 (N)	$\frac{TN}{TN+FN}$	$\frac{FN}{TN+FN}$	1
	1 (P)	$\frac{FP}{FP+TP}$	Precision : $\frac{TP}{FP+TP}$	1

Note : taille de la population totale et nombre de résultats de tests (positifs ou négatifs) ne sont plus lisibles avec cette normalisation.

Matrice confusion avec normalisation globale (éléments somment à 1)

		Classe réelle		
		0 (N)	1 (P)	
Classe prédite	0 (N)	$\frac{TN}{TP+FP+TN+FN}$	$\frac{FN}{TP+FP+TN+FN}$	
	1 (P)	$\frac{FP}{TP+FP+TN+FN}$	$\frac{TP}{TP+FP+TN+FN}$	
	$\sum_{/\text{ diag}}=$ taux d'erreur 1-accuracy			$\sum_{\backslash\text{diag}}=$ accuracy

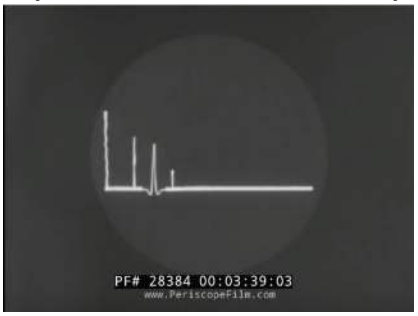
Note : La somme de tous les éléments dans la matrice est 1.

Receiver-operator characteristic curve (ROC) : origines

Radar Receiver

Deux "Receiver-operators"

Signal brut (ami ou ennemi ?)



U.S. NAVY WWII Radar Movie "Conquest of the Night"
`https://youtu.be/-BiBg2eOT-I?t=59`
`https://youtu.be/-BiBg2eOT-I?t=54`
`https://youtu.be/-BiBg2eOT-I?t=153`
La décision binaire (par exemple, ami/ennemi) est souvent prise sur la base d'une fonction de décision qui est seuillée. Comment choisir ce seuillage ?

Définition ROC (résumé des notations)

ROC :

$$\begin{aligned}\text{ROC} &= \text{Rappel en fonction de l'anti-specificité} \\ &= \text{TPR en fonction du FPR}\end{aligned}$$

Rappel = Sensitivité = True positive Rate (TPR) :

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

Spécificité = True Negative Rate (TNR)

$$\text{Spécificité} = \text{TNR} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

Anti-Spécificité = False Positive Rate (FPR)

$$\text{Anti-spécificité} = \text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = 1 - \text{TNR}$$

Définition ROC (suite)

Définition 3.11 (Courbe ROC) On appelle *courbe ROC* de l'anglais *Receiver-Operator Characteristic* la courbe décrivant l'évolution de la sensibilité (=rappel=TPR) d'un classifieur en fonction de la complémentaire à 1 de sa spécificité (=1-TNR = anti-specificité = FPR) obtenue en faisant varier le seuil.

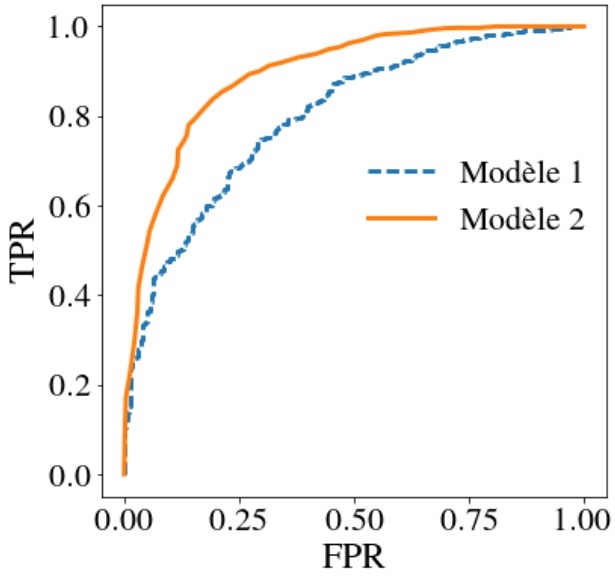


FIGURE 3.2 – Les courbes ROC de deux modèles.

Azencott

- (0,0) seuil haut, tel que toutes les étiquettes sont négatives
- (1,1) seuil bas, tel que toutes les étiquettes sont positives
- Chaque choix de seuil engendre une matrice de confusion !
- Classifieur idéal (aucune erreur) : passe par le point (0,1)
- Classifieur aléatoire : diagonale (0,0)–(1,1)

Exemple : construction d'une courbe ROC

Index de l'objet	1	2	3	4	5	6		
Etiquette (vraie, à trouver)	+	−	+	+	−	−		
Score retourné par le modèle	0.9	0.8	0.6	0.4	0.3	0.1	#(TP)	#(FP)
Prédiction si seuil > 0.9	FN	TN	FN	FN	TN	TN	0	0
Prédiction si $0.9 \geq$ seuil > 0.8	TP	TN	FN	FN	TN	TN	1	0
Prédiction si $0.8 \geq$ seuil > 0.6	TP	FP	FN	FN	TN	TN	1	1
Prédiction si $0.6 \geq$ seuil > 0.4	TP	FP	TP	FN	TN	TN	2	1
Prédiction si $0.4 \geq$ seuil > 0.3	TP	FP	TP	TP	TN	TN	3	1
Prédiction si $0.3 \geq$ seuil > 0.1	TP	FP	TP	TP	FP	TN	3	2
Prédiction si seuil ≤ 0.1	TP	FP	TP	TP	FP	FP	3	3

Si le score retourné par le modèle est plus grand ou égal au seuil, l'objet sera classifié comme positif et négatif sinon (puis TP,FP,TN ou FN en comparant avec l'étiquette)

Rappel = TPR = #(TP)/(#TP+#FN) = #(TP)/#(P) = TP/P
anti-spécificité = 1-TNR = FPR = #(FP)/(#FP+#TN) = #(FP)/#(N) = FP/N

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/N	0	0	1/3	1/3	1/3	2/3	1

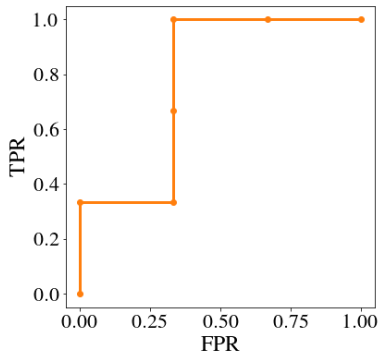


FIGURE 3.3 – Courbe ROC correspondant à l'expérience du tableau 3.2.

Azencott

Choix du seuil en fonction de la sensibilité ou spécificité qu'on souhaite garantir.

Courbe précision-rappel

Définition 3.12 (Courbe précision-rappel) on appelle *courbe précision-rappel* ou Precision-recall curve, la courbe décrivant l'évolution de la précision en fonction du rappel, lorsque le seuil de décision change.

Pour résumer l'aspect de cette courbe par un seul nombre, on peut utiliser l'aire sous celle-ci (appelée *area under the precision-recall curve (AUPR)*)

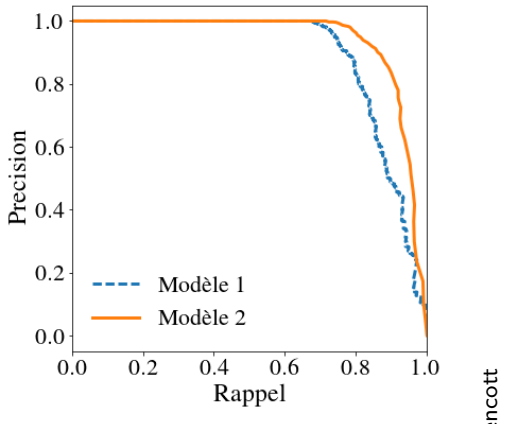


FIGURE 3.4 – Les courbes précision-rappel de deux modèles. Azencott

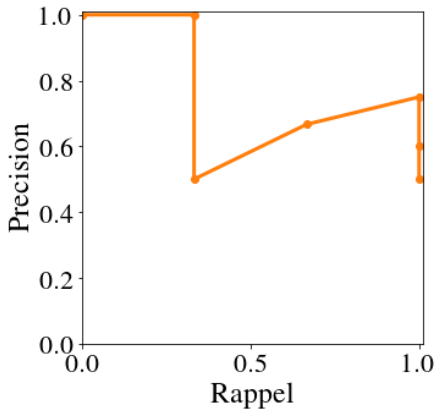


FIGURE 3.5 – Courbe précision-rappel correspondant à l'expérience du tableau 3.2. Azencott

Erreurs de régression

Pour les problèmes de régression, le nombre d'erreurs n'est pas un bon critère !

Idée : mesurer l'écart entre les prédictions et les valeurs réelles.

Définition 3.13 (Erreur quadratique moyenne (MSE)) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f(\vec{x}^1), f(\vec{x}^2), \cdots, f(\vec{x}^n)$, on appelle *erreur quadratique moyenne* ou mean squared error (MSE) la valeur

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

Définition 3.14 (RMSE) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f(\vec{x}^1), f(\vec{x}^2), \cdots, f(\vec{x}^n)$, on appelle *racine de l'erreur quadratique moyenne* ou root mean squared error (RMSE) la valeur [qui a la même unité que l'unité cible]

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

RMSLE (si couverture de plusieurs ordres de grandeur)

Définition 3.15 (RMSLE) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f\left(\vec{x}^1\right), f\left(\vec{x}^2\right), \cdots, f\left(\vec{x}^n\right)$, on appelle *racine du log de l'erreur quadratique moyenne* ou root mean squared log error (RMSLE) la valeur

$$\text{RMSLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\left(f\left(\vec{x}^i\right)+1\right) - \log\left(y_1^i+1\right)\right)^2}$$

Coefficient de détermination

Définition 3.16 (Coefficient de détermination) Étant données n étiquettes réelles $y^1, y^2, \ldots, y^n$ et n prédictions $f\left(\vec{x}^1\right)$, $f\left(\vec{x}^2\right)$, …, $f\left(\vec{x}^n\right)$, on appelle *erreur carrée relative* ou *relative squared error (RSE)* la valeur :

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

Le *coefficient de détermination* $R^2 = 1 - \text{RSE}$ est le carré du coefficient de corrélation entre $\vec{y}$ et $f\left(\vec{x}^1\right), f\left(\vec{x}^2\right), \ldots, f\left(\vec{x}^n\right)$: †

$$R = \frac{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{\ell=1}^{n} y^\ell\right)\left(f\left(\vec{x}^i\right) - \frac{1}{n}\sum_{\ell=1}^{n} f\left(\vec{x}^\ell\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(y^i - \frac{1}{n}\sum_{\ell=1}^{n} y^\ell\right)^2}\sqrt{\sum_{i=1}^{n}\left(f\left(\vec{x}^i\right) - \frac{1}{n}\sum_{\ell=1}^{n} f\left(\vec{x}^\ell\right)\right)^2}}$$

Ce coefficient indique à quel point les valeurs prédites sont corrélées aux valeurs réelles ; attention, il sera élevé aussi si elles leur sont anti-corrélées. [† voir note]

Méthodes naïves

Méthodes naïves pour la classification :

- prédire systématiquement l'étiquette majoritaire dans le jeu d'entrainement
- prédire une étiquette aléatoire
- prédire les scores de manière uniforme (classification binaire)

Méthodes naïves pour la régression :

- prédire une valeur aléatoire
- prédire systématiquement la médiane des étiquettes

Exemple d'utilisation : étant donné un indicateur quantitatif de performance (MSE, rappel, précision, etc.) et une méthode (pas naïve) qu'on veut caractériser :

1. évaluer la performance sur la méthode naïve, puis
2. évaluer la performance sur la méthode pas naïve

permettra de mettre en perspective la valeur numérique de la méthode pas naïve en comparaison de la méthode naïve (étalon).

Résumé cours 11

- La complexité du modèle engendre un dilemne (compromis) biais-variance
- 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.

Note concernant le coefficient de détermination (1)

Si les y^i sont les réalisations d'une variable aléatoire y, et $f^i = f(\vec{x}^i)$ les estimations de y^i telles qu'on a

$$y^i = f^i + e^i$$

avec e^i la réalisation d'une variable aléatoire de moyenne nulle qui n'est pas corrélée avec f^i, i.e. on a :

$$\mathrm{Cov}(f, e) = 0,$$

on peut alors réécrire

$$\mathrm{RSE} = \frac{\mathrm{Var}(e)}{\mathrm{Var}(y)}$$

et

Note concernant le coefficient de détermination (2)

$$\begin{aligned} R^2 &= \frac{(\mathrm{Cov}(y, f))^2}{\mathrm{Var}(y)\mathrm{Var}(f)} && (1)\\ &= \frac{(\mathrm{Cov}(f + e, f))^2}{\mathrm{Var}(y)\mathrm{Var}(f)} && (2)\\ &= \frac{(\mathrm{Cov}(f, f) + \mathrm{Cov}(e, f))^2}{\mathrm{Var}(y)\mathrm{Var}(f)} && (3)\\ &= \frac{(\mathrm{Var}(f))^2}{\mathrm{Var}(y)\mathrm{Var}(f)} && (4)\\ &= \frac{\mathrm{Var}(f)}{\mathrm{Var}(y)}. && (5) \end{aligned}$$

Note concernant le coefficient de détermination (3)

D'autre part

$$
\begin{aligned}
1 - \mathrm{RSE} &= \frac{\mathrm{Var}(y) - \mathrm{Var}(e)}{\mathrm{Var}(y)} && (6)\\
&= \frac{\mathrm{Var}(f)(\mathrm{Var}(y) - \mathrm{Var}(e))}{\mathrm{Var}(y)\mathrm{Var}(f)} && (7)\\
&= \frac{\mathrm{Var}(f)(\mathrm{Var}(f+e) - \mathrm{Var}(e))}{\mathrm{Var}(y)\mathrm{Var}(f)} && (8)\\
&= \frac{\mathrm{Var}(f)(\mathrm{Var}(f) + \mathrm{Var}(e) - \mathrm{Var}(e))}{\mathrm{Var}(y)\mathrm{Var}(f)} && (9)\\
&= \frac{\mathrm{Var}(f)(\mathrm{Var}(f)}{\mathrm{Var}(y)\mathrm{Var}(f)} && (10)\\
&= \frac{\mathrm{Var}(f)}{\mathrm{Var}(y)}. && (11)
\end{aligned}
$$