Machine Learning

Fiche de TD

Master BI&A

Guillaume Metzler
Institut de Communication (ICOM)
Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France
guillaume.metzler@univ-lyon2.fr

Résumé

Vingt-deux exercices : notions générales (biais et variance, sélection de modèle, fuite de données), deux exercices pour chaque algorithme (arbres de décision, forêts aléatoires, k plus proches voisins, régression logistique, SVM linéaire, SVM à noyau, AdaBoost, gradient boosting), l'évaluation d'un classifieur et quatre exercices sur l'apprentissage déséquilibré. Ils mêlent questions de compréhension, petits calculs faisables à la main (une calculatrice aide) et questions graphiques (lire ou tracer une figure). L'ensemble des étiquettes est précisé dans chaque exercice.

1 Notions générales

Exercice 1 – Complexité, diagnostic et sélection de modèle

1. Pour chaque hyperparamètre ci-dessous, dire si l'*augmenter* rend le modèle plus complexe (biais plus faible, variance plus forte, risque de sur-apprentissage) ou plus simple : le nombre de voisins k des kNN ; la profondeur maximale d'un arbre ; le nombre minimal d'exemples par feuille ; le nombre d'arbres T d'une forêt aléatoire ; la constante C d'un SVM à marge souple ; le paramètre γ du noyau gaussien $K(x, z) = e^{-\gamma\|x-z\|^2}$; la pénalité λ d'une régression logistique régularisée ; le nombre d'itérations M du boosting.

2. Trois modèles ont les erreurs suivantes (la meilleure erreur atteignable sur ce problème est d'environ 0.05). Diagnostiquer chacun (sous-apprentissage, sur-apprentissage, correct) et proposer un remède.

	modèle a	modèle b	modèle c
erreur d'apprentissage	0.01	0.21	0.07
erreur de validation	0.24	0.23	0.09

3. Un SVM à noyau RBF est réglé par recherche sur grille sur 4 valeurs de C et 5 valeurs de γ, avec une validation croisée à 5 plis, puis ré-entraîné sur tout l'ensemble d'apprentissage. Combien de SVM sont entraînés ? Pourquoi n'utilise-t-on pas l'ensemble de test pour choisir (C, γ) ?

4. Un étudiant standardise tout le jeu de données (moyenne et écart-type calculés sur tous les exemples), sélectionne ensuite les 20 variables les plus corrélées à l'étiquette, puis lance une validation croisée. Expliquer pourquoi l'erreur de validation croisée est optimiste, et comment corriger la démarche.

2 Arbres, forêts et plus proches voisins

Exercice 2 – Un arbre de décision à la main

Huit exemples décrits par deux variables, avec une étiquette $y \in \{0, 1\}$:

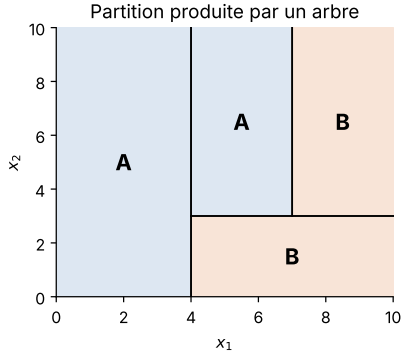
x_1	1	2	3	4	5	6	7	8
x_2	3	7	5	2	4	8	1	6
y	0	0	0	0	1	1	0	1

L'arbre utilise l'impureté de Gini $G = 2p(1-p)$ (p : proportion de la classe 1 dans le nœud) et choisit la coupure de plus forte diminution $G_{\text{nœud}} - \frac{n_G}{n_{\text{nœud}}} G_G - \frac{n_D}{n_{\text{nœud}}} G_D$ (G : gauche, D : droite).

1. Calculer l'impureté de Gini de la racine.
2. Calculer la diminution des quatre coupures candidates $x_1 < 4.5$, $x_1 < 6.5$, $x_2 < 3.5$, $x_2 < 7.5$. (Parmi toutes les coupures possibles, la meilleure est l'une de ces quatre.) Quelle coupure est choisie à la racine ?
3. L'un des fils est pur. Couper l'autre en choisissant entre $x_2 < 2.5$, $x_2 < 5$ et $x_2 < 7$. Dessiner l'arbre obtenu et la partition du plan correspondante. Quelle est son erreur d'apprentissage ?
4. Prédire la classe des points $(3, 8)$, $(6, 5)$ et $(7, 2)$.
5. Un arbre développé jusqu'au bout atteint toujours une erreur d'apprentissage nulle (si deux points identiques n'ont pas des étiquettes différentes). Pourquoi est-ce un problème ? Donner trois hyperparamètres qui le limitent. Un arbre a-t-il besoin de variables standardisées ? Pourquoi ?

Exercice 3 – Lire, construire et régler un arbre

1. **Question graphique.** La figure ci-dessous montre la partition du plan produite par un arbre de classification (classes A et B).

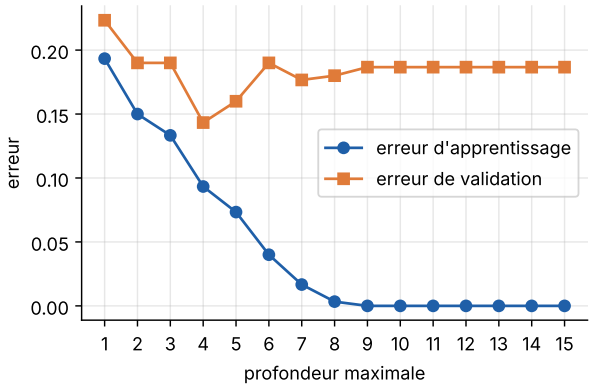


(a) Dessiner l'arbre correspondant (en commençant par la coupure de la racine). Combien a-t-il de feuilles, quelle est sa profondeur ? Prédire la classe de $(5, 5)$ et de $(8, 2)$.

(b) La vraie frontière entre les classes est en fait la diagonale $x_2 = x_1$. Esquisser sur la figure la façon dont un arbre l'approche. Pourquoi faut-il beaucoup de coupures ? Quel modèle de cette fiche serait mieux adapté ?

2. **Un arbre de régression.** Données $x = (1, 2, 3, 4, 5, 6)$, $y = (5, 7, 6, 12, 14, 13)$. Un arbre de régression prédit la moyenne de y dans chaque feuille et choisit la coupure qui minimise la somme des carrés des erreurs (SCE). Calculer la SCE de la racine, puis celle des coupures $x < 2.5$, $x < 3.5$ et $x < 4.5$. Laquelle est choisie, et quelles sont les valeurs des feuilles ?

3. **Question graphique.** Les erreurs d'apprentissage et de validation d'un arbre de classification sont tracées en fonction de sa profondeur maximale.



Quelle profondeur choisissez-vous ? Décrire les régimes à gauche et à droite de cette valeur. Pourquoi l'erreur d'apprentissage atteint-elle 0 ?

Exercice 4 – Forêts aléatoires : pourquoi et comment elles fonctionnent

1. Quelles sont les deux sources d'aléa d'une forêt aléatoire ? Pourquoi une seule ne suffit-elle pas ?

2. Chaque arbre a une variance de prédiction $\sigma^2 = 1$ et deux arbres ont une corrélation $\rho = 0.3$. La variance de la moyenne de T arbres vaut $\rho\sigma^2 + \frac{1-\rho}{T}\sigma^2$. La calculer pour $T = 1$, 10, 100 et $T \to \infty$. Que concluez-vous sur le rôle de T et de ρ ?

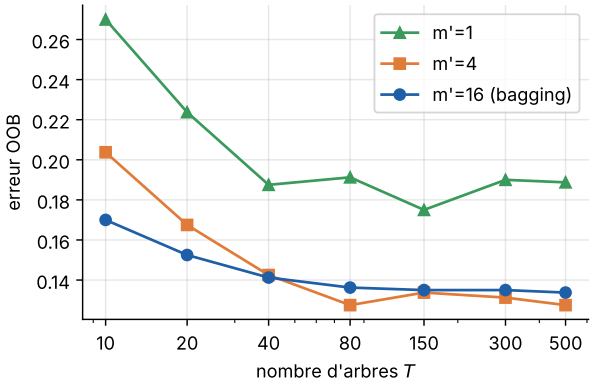
3. Il y a $d = 16$ variables, et à chaque coupure $m' = \sqrt{d} = 4$ variables sont tirées au hasard. Quelle est la probabilité qu'une variable donnée ne soit pas parmi les candidates ? Si seulement 2 variables sont informatives, quelle est la probabilité qu'aucune ne soit candidate ? Quel est l'effet de m' sur ρ et sur la qualité des arbres ?

4. Pourquoi les arbres d'une forêt sont-ils développés en profondeur (sans élagage) ? Une forêt peut-elle sur-apprendre quand T augmente ?

Exercice 5 – Forêts aléatoires en pratique

1. Une forêt de régression comporte trois arbres : h_1 : si $x_1 < 5$ alors 10 sinon 20 ; h_2 : si $x_2 < 3$ alors 12 sinon 18 ; h_3 : si $x_1 < 3$ alors 8, sinon (si $x_2 < 6$ alors 15 sinon 25). Calculer la prédiction de la forêt en $(4, 2)$ et en $(6, 7)$. Dans une forêt de classification de 100 arbres, 63 arbres votent pour la classe 1 en un point x : que prédit la forêt, et quelle « probabilité » donne habituellement `predict_proba` ?

2. **Question graphique.** L'erreur out-of-bag (OOB) d'une forêt est tracée en fonction du nombre d'arbres, pour trois valeurs de m' (nombre de variables tirées à chaque coupure, $d = 16$ variables).



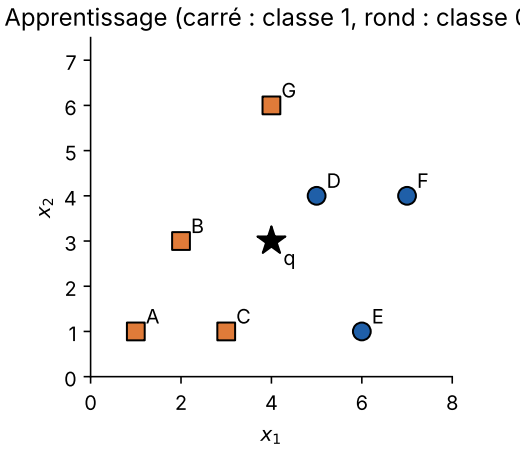
   Combien d'arbres suffisent ? Quel m' est le meilleur ici ? Expliquer le comportement de $m' = 1$ et de $m' = 16$.

3. Un data scientist ajoute comme variable le numéro de client (un entier différent pour chaque client). Son importance par diminution moyenne d'impureté (MDI) est la plus grande de toutes, mais son importance par permutation est quasi nulle. Expliquer. Que deviennent les importances quand une variable est dupliquée ?

Exercice 6 – k plus proches voisins

Sept points d'apprentissage dans le plan, d'étiquette dans $\{0, 1\}$, et un point requête $q = (4, 3)$; la distance est euclidienne.

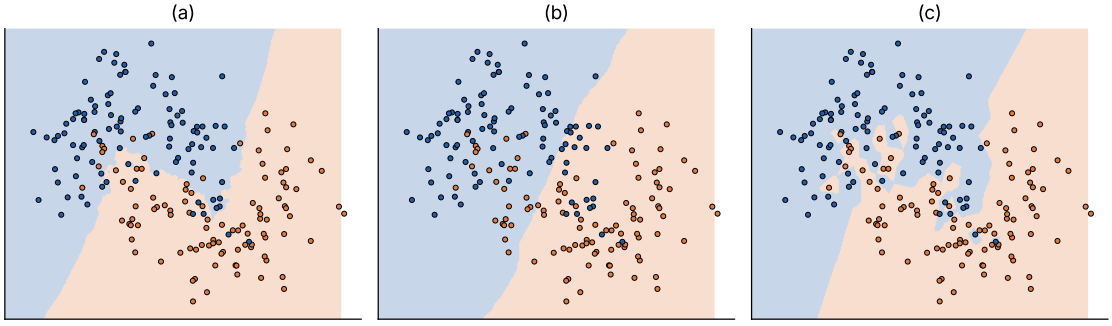
point	x_1	x_2	y
A	1	1	1
B	2	3	1
C	3	1	1
D	5	4	0
E	6	1	0
F	7	4	0
G	4	6	1



1. Calculer les distances au carré de q aux sept points et la prédiction des kNN (vote majoritaire) pour $k = 1$, 3 et 5.
2. La variable x_2 est maintenant exprimée dans une unité 10 fois plus petite (toutes les valeurs de x_2, y compris celle de q, sont multipliées par 10). Donner les nouvelles prédictions pour $k = 1$ et $k = 3$. Que concluez-vous ?
3. Pour choisir k, on utilise un jeu de données à une dimension : $x = (1, 3, 4, 6, 9, 10, 12, 13)$, d'étiquettes $y = (0, 1, 0, 0, 0, 1, 1, 1)$. Calculer l'erreur leave-one-out pour $k = 1$, 3 et 5 (chaque point est prédit par ses k plus proches voisins *parmi les autres* points). Quel k choisissez-vous ?
4. Que coûtent les kNN à l'apprentissage et à la prédiction (n exemples, d variables) ? Pourquoi fonctionnent-ils mal quand d est grand ?

Exercice 7 – Frontières, pondération et distances pour les kNN

1. **Question graphique.** Les trois figures montrent la frontière des kNN sur le même ensemble d'apprentissage pour $k \in \{1, 15, 150\}$ (200 points d'apprentissage), dans un ordre inconnu.



   Associer chaque figure à son k et à l'un des couples (erreur d'apprentissage, erreur de test) : (0, 0.145), (0.09, 0.075), (0.245, 0.22). Quel modèle généralise le mieux ?
2. **kNN en régression.** Points d'apprentissage $x = (1, 2, 4, 7, 9)$, $y = (3, 5, 4, 10, 12)$. Prédire y en $x = 5$ avec $k = 3$: (i) par la moyenne des voisins, (ii) avec des poids $\frac{1}{d}$ (inverse de la distance). Quel voisin compte le plus en (ii) ?
3. La requête est $q = (0, 0)$ et les deux points d'apprentissage sont $P_1 = (3, 0)$ et $P_2 = (2, 2)$. Lequel est le plus proche voisin pour la distance euclidienne ? Pour la distance de Manhattan $|u_1| + |u_2|$? Que concluez-vous ?

3 Modèles linéaires et SVM

Exercice 8 – Régression logistique : prédiction, perte et un pas de gradient

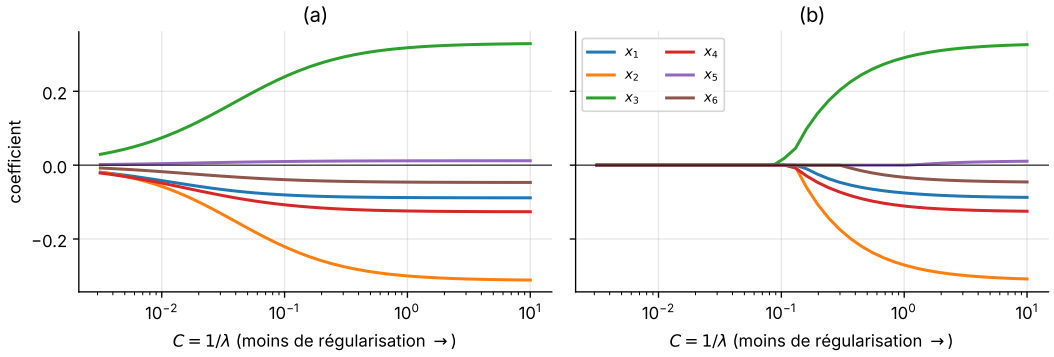
Étiquettes $y \in \{0, 1\}$, modèle $p(y = 1 \mid x) = \sigma(w_0 + w_1 x_1 + w_2 x_2)$ avec $\sigma(z) = \frac{1}{1+e^{-z}}$, poids actuels $(w_0, w_1, w_2) = (-3,\ 1,\ 0.5)$. Données :

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
(x_1, x_2)	$(1, 2)$	$(2, 4)$	$(4, 2)$	$(3, 0)$
y	0	1	1	0

1. Calculer le score z, la probabilité p et la classe prédite (seuil 0.5) de chaque exemple, puis la perte logistique moyenne $\frac{1}{n}\sum_i -[y_i \ln p_i + (1-y_i)\ln(1-p_i)]$.
2. Montrer que le gradient de cette perte par rapport à $w = (w_0, w_1, w_2)$ vaut $\frac{1}{n}\sum_i (p_i - y_i)\,\tilde{x}_i$ avec $\tilde{x}_i = (1, x_{i1}, x_{i2})$ (utiliser $\sigma' = \sigma(1-\sigma)$). Le calculer, puis faire un pas de gradient de taille $\eta = 0.5$. La perte diminue-t-elle ?
3. Interpréter $w_1 = 1$ et $w_2 = 0.5$ en termes de cote $\frac{p}{1-p}$ (*odds*). Donner l'équation de la frontière de décision.
4. Ces données sont linéairement séparables. Que devient $\|w\|$ si l'on minimise la perte sans régularisation ? Que change une pénalité L_2, $\lambda\|w\|^2$?

Exercice 9 – Régression logistique : plusieurs classes, variables non linéaires, régularisation

1. **Softmax.** Avec 3 classes, le modèle calcule un score $z_k = w_k \cdot x$ par classe et $p_k = \frac{e^{z_k}}{\sum_l e^{z_l}}$. Pour un exemple de scores $z = (2, 1, -1)$, calculer les probabilités et la classe prédite. La vraie classe est la classe 2 : calculer l'entropie croisée $-\ln p_2$. Que deviennent les probabilités si l'on ajoute la même constante aux trois scores ?
2. **Variables non linéaires.** Une régression logistique est entraînée sur les variables $(x_1^2,\ x_2^2)$: $p(y = 1 \mid x) = \sigma(-4 + x_1^2 + x_2^2)$. Donner l'équation et la forme de la frontière de décision, et les probabilités en $(0, 0)$, $(1, 1)$ et $(2, 2)$.
3. **Question graphique.** Les coefficients d'une régression logistique à 6 variables standardisées sont tracés en fonction de $C = \frac{1}{\lambda}$, avec une pénalité L_1 sur une figure et une pénalité L_2 sur l'autre.

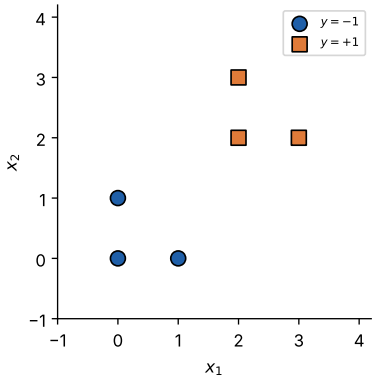


Quelle figure correspond à L_1 ? Sur cette figure, dans quel ordre les variables entrent-elles dans le modèle ? Quelles variables garderiez-vous ? À quoi ressemblent les deux figures quand C est grand, et pourquoi ?

Exercice 10 – SVM linéaire : marge, vecteurs supports et perte charnière

Étiquettes $y \in \{-1, +1\}$, classifieur $\text{sign}(w \cdot x + b)$. Données :

$y=-1$	$(0,0)$, $(1,0)$, $(0,1)$
$y=+1$	$(2,2)$, $(3,2)$, $(2,3)$



Le SVM à marge dure résout $\min \frac{1}{2}\|w\|^2$ sous les contraintes $y_i(w \cdot x_i + b) \geq 1$ pour tout i.

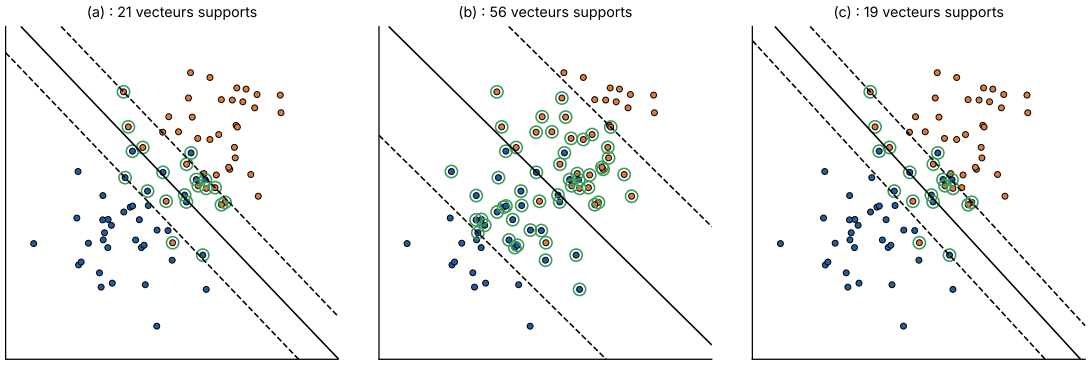
1. Par symétrie des données, justifier que w est proportionnel à $(1,1)$. Montrer que $w = (\frac{2}{3}, \frac{2}{3})$, $b = -\frac{5}{3}$ vérifie toutes les contraintes, avec égalité pour trois points. Quels sont les vecteurs supports ? Calculer la marge $\frac{2}{\|w\|}$ et la vérifier sur la figure.
2. La solution duale s'écrit $w = \sum_i \alpha_i y_i x_i$ avec $\alpha_i \geq 0$, $\sum_i \alpha_i y_i = 0$, et $\alpha_i > 0$ seulement pour les vecteurs supports. Trouver les α_i.
3. Classer le nouveau point $(3,0)$. Est-il dans la marge ?
4. On ajoute un nouvel exemple $(2,1)$ d'étiquette -1. Calculer sa perte charnière (*hinge*) $\max(0, 1 - y(w \cdot x + b))$ pour le (w, b) actuel. Les données sont-elles encore linéairement séparables ? Pour un SVM à marge souple $\min \frac{1}{2}\|w\|^2 + C\sum_i \max(0, 1 - y_i(w \cdot x_i + b))$, que se passe-t-il quand C est très grand ou très petit ?

Exercice 11 – SVM à marge souple : variables d'écart et constante C

1. On considère le classifieur linéaire $f(x) = x_1 - 3$ (soit $w = (1,0)$, $b = -3$) et les exemples suivants (seul x_1 compte) :

exemple	A	B	C	D	E	F	G
x_1	5	4	3.5	2.5	1	2	3.4
y	$+1$	$+1$	$+1$	$+1$	-1	-1	-1

   Calculer $y_i f(x_i)$ et la variable d'écart $\xi_i = \max(0, 1 - y_i f(x_i))$ de chaque exemple. Les classer : hors de la marge, sur la marge, dans la marge mais bien classé, mal classé. Calculer l'objectif $\frac{1}{2}\|w\|^2 + C\sum_i \xi_i$ pour $C = 1$ et $C = 10$.
2. Si ce f était la solution du SVM, quels exemples auraient $\alpha_i = 0$, $0 < \alpha_i < C$, $\alpha_i = C$? Lesquels sont des vecteurs supports ?
3. **Question graphique.** Un SVM linéaire est entraîné sur les mêmes données avec $C = 0.01$, $C = 1$ et $C = 100$ (trait plein : frontière, pointillés : marge, cerclés : vecteurs supports), dans un ordre inconnu.



Associer chaque figure à sa valeur de C. Expliquer le nombre de vecteurs supports.

4. Pourquoi faut-il standardiser les variables avant un SVM ? Avec $K = 4$ classes, combien de SVM binaires entraîne-t-on en un-contre-tous et en un-contre-un ?

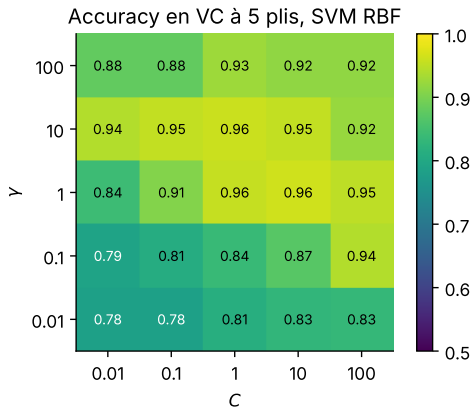
Exercice 12 – SVM à noyau sur le problème du XOR

Quatre points d'étiquettes dans $\{-1, +1\}$: $(1, 1)$ et $(-1, -1)$ ont l'étiquette $+1$; $(1, -1)$ et $(-1, 1)$ ont l'étiquette -1.

1. Montrer qu'aucun classifieur linéaire $\text{sign}(w \cdot x + b)$ ne sépare ces points (additionner les scores des deux points positifs, puis ceux des deux négatifs).
2. On considère le noyau polynomial $K(x, z) = (x \cdot z)^2$. Vérifier que $K(x, z) = \phi(x) \cdot \phi(z)$ avec $\phi(x) = (x_1^2,\ \sqrt{2}\, x_1 x_2,\ x_2^2)$. Calculer la matrice de Gram 4×4. Représenter les points dans l'espace de redescription : y sont-ils séparables ?
3. Le SVM à noyau classe avec $f(x) = \sum_i \alpha_i y_i K(x_i, x) + b$. Par symétrie, tous les α_i sont égaux à un même α et $b = 0$. Montrer que $f(x) = 8\alpha\, x_1 x_2$. Trouver α pour que les quatre points vérifient $y_i f(x_i) = 1$ (ce sont tous des vecteurs supports). Classer $(2, 0.5)$ et $(0.5, -3)$ et décrire les régions de décision.
4. Avec le noyau gaussien $K(x, z) = e^{-\gamma \|x - z\|^2}$, calculer K entre $(1, 1)$ et son voisin $(1, -1)$ ($\|x - z\|^2 = 4$) et le point opposé $(-1, -1)$ ($\|x - z\|^2 = 8$), pour $\gamma = 0.5$ et $\gamma = 2$. Que deviennent la matrice de Gram et le classifieur quand γ est très grand ?

Exercice 13 – Noyaux : calculs, validité et réglage

1. Calculer la matrice de Gram du noyau gaussien $K(x, z) = e^{-\gamma (x - z)^2}$, $\gamma = 0.5$, pour les trois réels $x = 0$, 1, 3. Quels points sont « semblables » ?
2. Pour $x, z \in \mathbb{R}^2$, développer $K(x, z) = (x \cdot z + 1)^2$ et donner une fonction de redescription ϕ telle que $K(x, z) = \phi(x) \cdot \phi(z)$. Calculer $K((1, 2), (3, -1))$ directement et avec ϕ. Avec $d = 100$ variables et le noyau $(x \cdot z + 1)^3$, l'espace de redescription est de dimension $\binom{d+3}{3}$: la calculer. Qu'économise l'astuce du noyau ?
3. $K(x, z) = -\|x - z\|^2$ est-il un noyau valide ? (Indication : matrice de Gram de deux points à distance 1.)
4. **Question graphique.** Le taux de bonne classification en validation croisée d'un SVM à noyau RBF sur un jeu de données non linéaire est donné pour une grille de (C, γ).



Quel couple choisissez-vous ? Expliquer les faibles performances dans le coin inférieur gauche et la baisse pour $\gamma = 100$.

4 Boosting

Exercice 14 – AdaBoost sur trois tours

Six points sur une droite, d'étiquettes dans $\{-1, +1\}$:

x	1	2	3	4	5	6
y	$+1$	$+1$	-1	-1	$+1$	-1

Les classifieurs faibles sont les souches (*stumps*) $h_{\theta,s}(x) = s \cdot \mathrm{sign}(x-\theta)$ avec $\theta \in \{0.5, 1.5, \ldots, 6.5\}$ et $s \in \{-1, +1\}$ ($s = -1$ prédit $+1$ à gauche de θ). AdaBoost part de poids uniformes $w_i = \frac{1}{6}$; à chaque tour, il choisit la souche d'erreur pondérée ε_t minimale, pose $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$ et met à jour les poids : les points mal classés sont multipliés par $\frac{1}{2\varepsilon_t}$ et les autres par $\frac{1}{2(1-\varepsilon_t)}$ (c'est la mise à jour $w_i \, e^{-\alpha_t y_i h_t(x_i)}$ suivie de la normalisation).

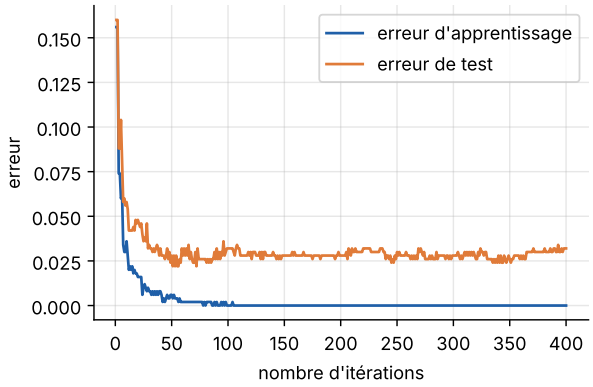
1. **Tour 1.** Trouver la meilleure souche, ε_1, α_1 et les nouveaux poids.
2. **Tour 2.** Avec ces poids, trouver la meilleure souche (elle se trompe sur deux points), ε_2, α_2 et les nouveaux poids. Calculer $F_2 = \alpha_1 h_1 + \alpha_2 h_2$ sur les six points et l'erreur d'apprentissage de $\mathrm{sign}(F_2)$.
3. **Tour 3.** Idem : meilleure souche, ε_3, α_3, puis F_3 et son erreur d'apprentissage.
4. L'erreur d'apprentissage est majorée par $\prod_t 2\sqrt{\varepsilon_t(1-\varepsilon_t)}$. Calculer cette borne après trois tours. Pourquoi AdaBoost augmente-t-il le poids des points mal classés ? Quel est son point faible avec des étiquettes bruitées ?

Exercice 15 – AdaBoost et la perte exponentielle

1. **Question graphique.** Avec la marge $m = yF(x)$, tracer sur une même figure, pour $m \in [-2, 2]$: la perte 0-1 $\mathbb{1}[m \leq 0]$, la perte exponentielle e^{-m} (AdaBoost), la perte logistique $\log_2(1 + e^{-m})$ (régression logistique, gradient boosting) et la perte charnière $\max(0, 1 - m)$ (SVM). Les calculer pour $m = -1, 0, 1, 2$. Quelle perte pénalise le plus un point très mal classé ?

2. Après quelques tours, l'ensemble F donne les marges $y_i F(x_i) = (2,\ 0.5,\ -0.3,\ 1,\ -1.5)$ sur cinq exemples. Les poids d'AdaBoost sont $w_i \propto e^{-y_i F(x_i)}$. Calculer les poids normalisés. Sur quels exemples le prochain classifieur faible va-t-il se concentrer ?

3. **Question graphique.** Les erreurs d'apprentissage et de test d'AdaBoost avec des souches sont tracées en fonction du nombre d'itérations.



Lire les erreurs après 1, 10 et 100 itérations. L'erreur d'apprentissage est nulle après environ 100 itérations : pourquoi l'erreur de test n'augmente-t-elle pas ensuite ? Comment choisiriez-vous le nombre d'itérations ?

Exercice 16 – Gradient boosting en régression avec des souches

Données : $x = (1, 2, 3, 4, 5)$, $y = (1, 2, 3, 10, 14)$, perte quadratique $\frac{1}{2}(y - F)^2$. On part de $F_0 = \bar{y}$. À chaque itération, une souche de régression (coupure $x < \theta$, $\theta \in \{1.5, 2.5, 3.5, 4.5\}$, chaque feuille prédisant la moyenne de ses résidus) est ajustée aux résidus $r_i = y_i - F_{t-1}(x_i)$ en minimisant la somme des carrés des erreurs (SCE), et $F_t = F_{t-1} + \nu h_t$ avec $\nu = 0.5$.

1. Montrer que pour la perte quadratique les pseudo-résidus $-\frac{\partial \ell}{\partial F}$ sont les résidus $y - F$. Calculer F_0, son erreur quadratique moyenne (EQM) et les résidus.

2. Itération 1 : calculer la SCE des quatre coupures, en choisir une, donner les valeurs des feuilles, F_1 et son EQM.

3. Itération 2 : idem avec les nouveaux résidus. La coupure est-elle la même qu'à l'itération 1 ? Donner F_2 et son EQM.

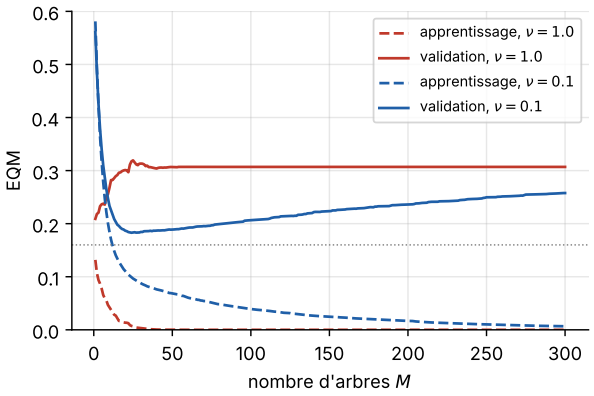
4. Quel est le rôle de ν et du nombre d'itérations ? Comment les choisit-on ? Que devient AdaBoost dans ce cadre ?

Exercice 17 – Gradient boosting : prédiction, taux d'apprentissage et comparaison avec les forêts

1. Un modèle de gradient boosting en régression a $F_0 = 50$, un taux d'apprentissage $\nu = 0.1$ et trois arbres : h_1 : si $x < 10$ alors -20 sinon 30 ; h_2 : si $x < 5$ alors -15 sinon 10 ; h_3 : si $x < 20$ alors 5 sinon -8. Prédire pour $x = 8$ et $x = 25$.

2. Avec la perte absolue $|y - F|$, calculer les pseudo-résidus pour les résidus $y - F = (-3,\ 0.5,\ 8)$. Pourquoi cette perte est-elle plus robuste aux valeurs aberrantes ?

3. **Question graphique.** Les EQM d'apprentissage (tirets) et de validation (traits pleins) sont tracées en fonction du nombre d'arbres, pour $\nu = 1$ et $\nu = 0.1$ (arbres de profondeur 3 ; la ligne pointillée est le niveau du bruit).



Pour chaque ν, lire le meilleur nombre d'arbres et la meilleure erreur de validation. Quel réglage préférez-vous ? Qu'est-ce que l'arrêt précoce ?

4. Comparer forêts aléatoires et gradient boosting : profondeur des arbres, construction des arbres (en parallèle ou séquentielle), ce qui est réduit (biais ou variance), effet du nombre d'arbres, sensibilité aux hyperparamètres.

5 Évaluation d'un classifieur

Exercice 18 – Évaluer un classifieur et choisir un modèle

Un test de dépistage est évalué sur 1000 patients, dont 50 sont malades (classe positive). Matrice de confusion d'un classifieur :

	prédit malade	prédit sain
malade	40	10
sain	90	860

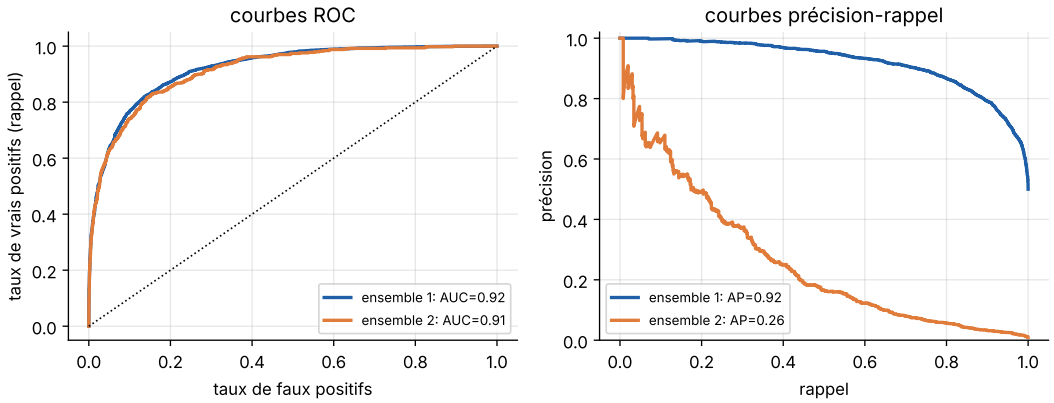
1. Calculer le taux de bonne classification (*accuracy*), la précision, le rappel (sensibilité), la spécificité, le score F_1 et l'accuracy équilibrée.
2. Quelle est l'accuracy du classifieur qui prédit toujours « sain » ? Que concluez-vous ?
3. Avec un seuil de décision plus élevé, la matrice de confusion devient (30, 20 | 30, 920) (même disposition). Calculer les mêmes métriques. Quel seuil préférez-vous pour un test de dépistage, et pourquoi ? Quel outil résume l'ensemble des seuils ?
4. Pour chaque situation, proposer un ou deux modèles adaptés parmi ceux de cette fiche, et justifier : (a) les médecins veulent des règles de décision explicites ; (b) 100 000 exemples, 30 variables tabulaires hétérogènes, la meilleure performance possible ; (c) 200 exemples et 10 000 variables (génomique) ; (d) on a besoin de probabilités bien calibrées ; (e) des millions d'exemples d'apprentissage et des prédictions rapides.

6 Apprentissage déséquilibré

Exercice 19 – Prévalence, précision et choix de la courbe

Un classifieur a un taux de vrais positifs (rappel) $TVP = 0.9$ et un taux de faux positifs $TFP = 0.05$, quel que soit l'ensemble de test. On note π la proportion de positifs (prévalence) dans la population.

1. Montrer que la précision vaut $\frac{TVP\,\pi}{TVP\,\pi + TFP\,(1-\pi)}$. La calculer pour $\pi = 0.5$, 0.1 et 0.01. Commenter.
2. **Question graphique.** Le même modèle de score est évalué sur deux ensembles de test : le premier contient 50 % de positifs, le second 1 %. À gauche : courbes ROC ; à droite : courbes précision-rappel (AP : précision moyenne, l'aire sous la courbe PR).



Pourquoi les courbes ROC sont-elles presque identiques alors que les courbes PR diffèrent autant ? Quelle est la précision d'un classifieur aléatoire sur chaque ensemble de test ? Quelle courbe présenteriez-vous pour un problème de détection de fraude, et pourquoi ?

Exercice 20 – Courbe ROC et AUC à la main

Scores donnés par un classifieur à dix exemples de test (P : positif, N : négatif), triés par score décroissant :

score	0.95	0.90	0.85	0.70	0.60	0.50	0.40	0.30	0.20	0.10
étiquette	P	N	P	N	P	N	P	N	N	N

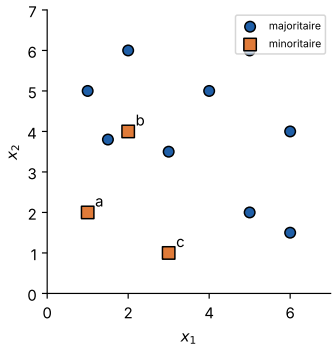
1. Pour chaque seuil t (un exemple est prédit positif si son score est $\geq t$), t prenant les dix valeurs des scores, calculer le TVP et le TFP. **Tracer** la courbe ROC.
2. Calculer l'AUC comme la proportion de paires (positif, négatif) dans lesquelles le positif a le score le plus élevé. La vérifier avec l'aire sous votre courbe.
3. Quel seuil garantit un rappel d'au moins 0.75 avec le moins de faux positifs ? Donner la précision correspondante.
4. Recalculer l'AUC si chaque score est remplacé par son carré. Que mesure l'AUC ?

Exercice 21 – Rééquilibrer l'ensemble d'apprentissage : sous-échantillonnage, sur-échantillonnage, SMOTE

Un ensemble d'apprentissage contient 2000 exemples, dont 100 positifs (classe minoritaire).

1. Donner le ratio de déséquilibre. Avec un sous-échantillonnage aléatoire de la classe majoritaire pour obtenir un ratio 1:1, combien d'exemples sont retirés ? Avec un sur-échantillonnage aléatoire (duplication) de la classe minoritaire, combien de copies sont ajoutées ? Donner un inconvénient de chacun.

2. SMOTE crée un exemple minoritaire synthétique $x_{\text{nouv}} = x + u\,(x_{pv} - x)$, $u \sim \mathcal{U}[0,1]$, où x_{pv} est l'un des plus proches voisins minoritaires de x. **Question graphique :** sur la figure, tracer les segments sur lesquels SMOTE peut créer des points à partir de $a = (1,2)$, $b = (2,4)$, $c = (3,1)$. Calculer les points synthétiques pour $(x, x_{pv}, u) = (a, b, 0.25)$ et $(a, c, 0.5)$.



3. D'après la figure, quel risque SMOTE prend-il avec le segment de b à c ?

4. Un étudiant applique SMOTE à tout le jeu de données, puis lance une validation croisée à 5 plis. Qu'est-ce qui ne va pas ? Faut-il rééquilibrer l'ensemble de test ?

Exercice 22 – Coûts, poids de classes et déplacement du seuil

Dans un problème de défaut de crédit, manquer un client défaillant (faux négatif) coûte $C_{FN} = 10$, refuser un bon client (faux positif) coûte $C_{FP} = 1$.

1. Un modèle calibré donne $p = P(y = 1 \mid x)$. Prédire positif coûte en moyenne $(1-p)C_{FP}$, prédire négatif coûte $p\,C_{FN}$. En déduire le seuil optimal $t^\star$. Comparer à 0.5.

2. Dix clients ont les probabilités et étiquettes ci-dessous. Calculer VP, FP, FN et le coût total $10\,FN + FP$ pour les seuils 0.5 et $t^\star$.

p	0.95	0.80	0.60	0.45	0.30	0.20	0.15	0.12	0.05	0.02
y	1	0	1	0	1	0	0	1	0	0

3. Au lieu de déplacer le seuil, on peut pondérer la perte par classe. Les poids « balanced » de scikit-learn sont $w_k = \frac{n}{K\,n_k}$. Les calculer pour $n = 1000$, $n_1 = 50$, $n_0 = 950$. Que changent-ils dans une régression logistique ou un SVM ?

4. Un modèle est entraîné après n'avoir gardé qu'une fraction $\beta = 0.1$ des négatifs (sous-échantillonnage). Ses probabilités p_s sont alors trop élevées ; on les corrige par $p = \frac{\beta p_s}{\beta p_s - p_s + 1}$. Corriger $p_s = 0.2$, 0.5 et 0.9. Pourquoi cette correction est-elle nécessaire ?