Introduction à l'apprentissage automatique

Tronc commun scientifique 2A - Mines Nancy

Examen - mardi 24 janvier 2023

Durée : 2 heures 30 minutes

Ce test comporte deux parties. À la fin du test, rendez votre copie (partie 1) au surveillant et déposez sur Arche votre carnet Jupyter (partie 2).

La qualité et la précision de la rédaction seront valorisées.

Note finale au module :

$$\text{Note de TD (4 points)} + \frac{16}{20}\big(\text{Note partie 1 (10 points)} + \text{Note partie 2 (10 points)}\big)$$

Partie 1

Sauf dans la question 1 (où on ne demande pas de justification), justifiez brièvement vos réponses (pas de point sans justification).

1) *0.5pt* Parmi les affirmations suivantes sur la validation croisée, laquelle ou lesquelles est/sont correcte(s) ?

- **a)** certaines observations peuvent participer à plusieurs blocs ou plis (*folds*) de test ;
- **b)** à la fin du processus de validation croisée, chaque bloc ou pli (*fold*) a été utilisé exactement une fois dans l'entraînement d'un seul modèle ;
- **c)** dans cette procédure, la validation croisée à K plis nécessite l'entraînement de K modèles ;
- **d)** la validation croisée permet de choisir des valeurs d'hyperparamètres.

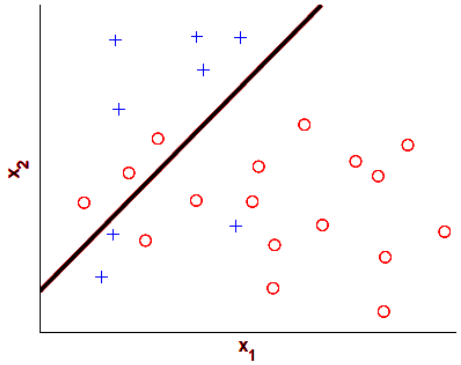
2) *0,5pt* On suppose disposer de dix observations $(x_i, y_i)_{1\leqslant i\leqslant 10}$ où les x_i et y_i sont des réels. On considère les deux modèles de régression suivants :

— *Modèle 1* : $y = w_0 + w_1 x$

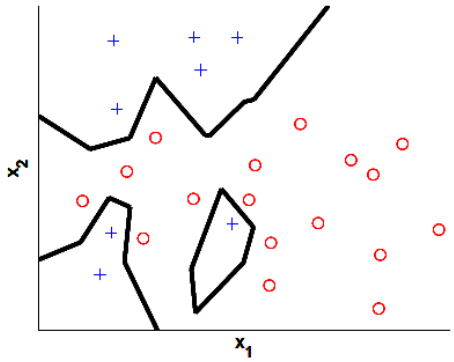
— *Modèle 2* : $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + w_4 x^4 + w_5 x^5 + w_6 x^6 + w_7 x^7 + w_8 x^8 + w_9 x^9$

Quel modèle va probablement présenter du surapprentissage ?

3) *1,5pt* Les figures suivantes représentent les observations d'apprentissage (classes + et o) dans un problème de classification supervisée d'observations dans $\mathbb{R}^2$, ainsi que les frontières de séparation obtenues par deux classifieurs.



Cas 1



Cas 2

Quels classifieurs vus en cours sont susceptibles d'avoir été utilisés dans chaque cas ? Quel est le nombre d'erreurs sur la base d'apprentissage ?

4) *3,5pt* Contrôle d'une démarche expérimentale.

Vous êtes responsable d'une équipe d'experts IA. Pour chaque situation, dites si elle est, selon vous, **correcte** ou **problématique**. Si la situation est problématique, justifiez brièvement et expliquez ce qu'il faudrait faire.

— Un de vos experts rapporte une erreur d'entraînement très faible et en déduit que le modèle utilisé est bon.

— Un des experts travaille sur un problème de détection de fraudes dans des données bancaires. La base d'apprentissage comprend 100 exemples de fraudes et 10000 exemples « sans fraude ». L'expert assure avoir obtenu un grand succès dans ce problème de classification supervisée, car il obtient 98% de classifications correctes sur des données de contrôle répartis de la même manière que la base d'apprentissage.

— Un des experts utilise un modèle de classification possédant deux hyperparamètres. Il sépare ses données entre apprentissage et test. En utilisant la procédure de validation croisée sur les données d'apprentissage, l'expert choisit les meilleurs hyperparamètres, puis entraîne le modèle correspondant sur la base d'apprentissage entière. Il vous rapporte ensuite les résultats sur la base de test.

— Un expert travaille sur la détection automatique de cancers de poumon dans des images-scanners, à l'aide de modèles de classification supervisée. On dispose de dix scanners de poumons par patient, chez des patients sains et malades. Chez les patients malades, les scanners sont pris alors que la maladie est déjà déclarée (les tumeurs sont visibles sur chaque scanner). L'intérêt d'utiliser dix scanners est de tenir compte de la variabilité d'acquisition (types de scanners différents, mouvements du patient) ou les différents stades de la maladie. L'étude regroupe les scanners de 50 patients sains et 50 patients malades. L'expert forme une base de données unique de toutes les images (1000 images donc), et fait une séparation aléatoire entre base d'apprentissage (80% des images) et base de test (20% des images). Le modèle est entraîné sur la base d'apprentissage et évalué sur la base de test.

5) *2pt* On considère un problème de classification dans lequel les observations (x_1, x_2) sont bidimensionnelles et les étiquettes y des deux classes sont +1 et -1. La base d'apprentissage est formée des quatres observations étiquetées suivantes :

x_1	x_2	y
0	0	-1
1	0	$+1$
0	1	$+1$
1	1	-1

Une SVM linéaire permet-elle de séparer les deux classes ?

On considère la fonction de plongement dans l'espace de redescription $\mathbb{R}^3$ suivante :

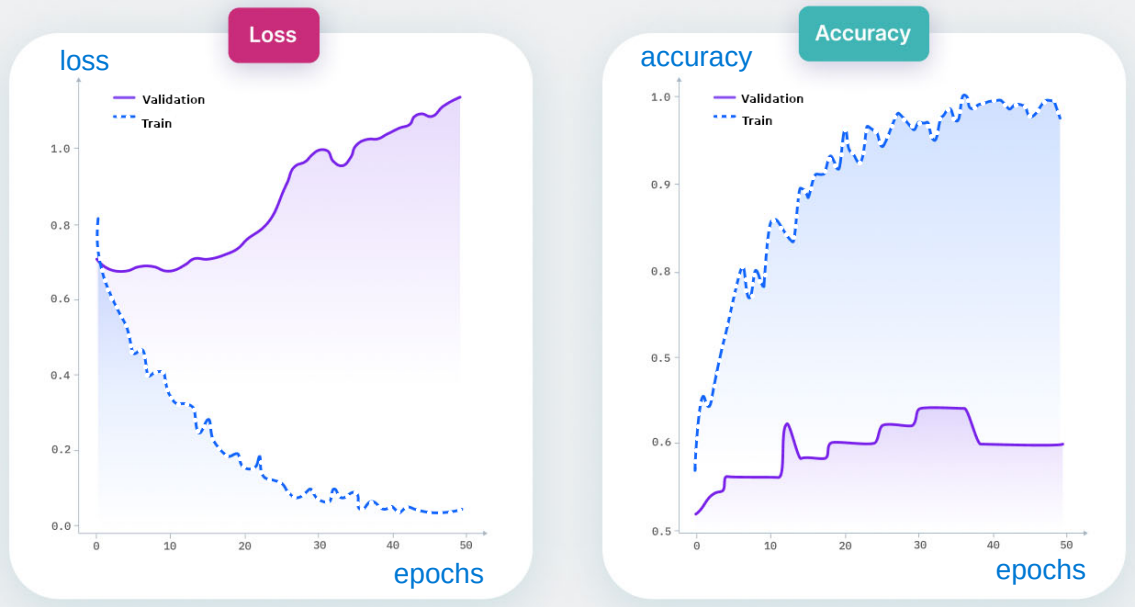
$$\phi(x_1, x_2) = \left(x_1^2 - x_2^2,\, x_1 x_2,\, x_1^2 + x_2^2\right)$$

À quel noyau k cette fonction ϕ correspond-elle ?

On admet que ce noyau satisfait les conditions de Mercer. Une SVM avec ce noyau permet-elle de classifier correctement la base d'apprentissage ?

Vous ferez des dessins dans $\mathbb{R}^2$ et $\mathbb{R}^3$.

6) *1,5pt* Lors de l'entraînement d'un réseau de neurones profond sur un problème de classification supervisée, on obtient les courbes suivantes.



Que représentent ces courbes ? Qu'est ce que désigne *accuracy* ici ?

Est-il logique que la courbe pointillée de la figure de gauche soit globalement décroissante ? Quel peut être l'explication des « oscillations » autour de la tendance décroissante ?

Quelle conclusion tire-t-on de ces graphiques sur le modèle ou la base d'apprentissage ?

7) *0,5pt* On dispose de la matrice D des distances d'édition entre cent mots (des chaînes de caractères). Autrement dit, l'élément de ligne i et colonne j de la matrice D est la distance entre les mots d'indices i et j ($1 \leqslant i, j \leqslant 100$). Quel(s) algorithme(s) de partitionnement vu(s) en cours ou TD pouvez-vous mettre en œuvre, quel(s) algorithme(s) ne pouvez-vous pas utiliser, et pourquoi ?

Partie 2

10pts Étude d'un problème pratique, disponible sur Arche.

Déposez votre carnet avec les cellules exécutées au format HTML ("`File`" → "`Download as`" → "`HTML`"). On doit pouvoir vous évaluer à partir de ce fichier HTML visualisé dans un navigateur. Préparez votre dépôt quelques minutes avant la fin du test pour ne pas dépasser la limite de temps.