Introduction à l'apprentissage automatique

Tronc commun scientifique 2A - Mines Nancy

Examen - mardi 16 janvier 2024

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).

Documents autorisés **(à l'exclusion de tout autre)** : notes et fichiers personnels, polycopié ; et sur internet : page arche du cours et documentation scikit-learn.

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

Justifiez toutes vos réponses (pas de point sans justification).

1) *1.5pt* On se place dans le cadre de l'apprentissage supervisé.

A. On considère un problème de régression. Pour des ensembles d'apprentissage et de test fixes, on procède à l'apprentissage de modèles polynomiaux de degré 1, puis 2, puis 3, etc. Représentez l'allure des courbes représentant l'erreur de prédiction sur la base d'apprentissage et sur la base test en fonction du degré.

B. On considère une base d'apprentissage de 5000 observations distribuées équitablement dans deux classes, les observations étant décrites par 10 caractéristiques. On remarque que le classifieur de la régression logistique souffre d'un fort sous-apprentissage. Par conséquent, quel(s) classifieur(s) essayez-vous parmi les classifieurs suivants :

(a) une SVM linéaire ;

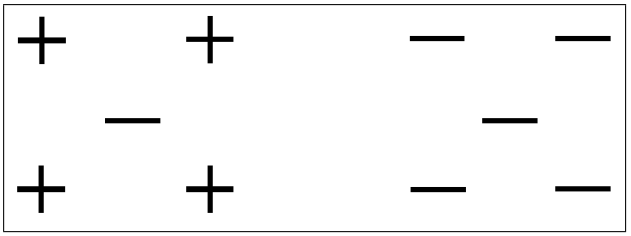
(b) une SVM à noyau RBF (gaussien) ;

(c) le perceptron de Rosenblatt ;

(d) un perceptron multicouche dont les fonctions d'activations sont définies par $f(x) = \alpha x$, où α est un paramètre réel.

C. On travaille avec une base d'apprentissage de 1000 observations distribuées dans trois classes A, B, C, de la manière suivante : 800 observations sont dans la classe A, 150 dans B, et 50 dans C. Un camarade vous conseille d'utiliser XGBoost, qui vous fournit un classifieur de taux moyen de classifications correctes de 70% sur une base de test de même distribution statistique que la base d'apprentissage. Qu'en pensez-vous ?

2) *1,5pt* Pour le jeu de données représenté dans l'encadré ci-dessous (observations dans $\mathbb{R}^2$, étiquettes **+** ou **-**), calculez le score de validation croisée *leave-one-out* des classifieurs K plus proches voisins pour $K = 1$ et $K = 3$. Dans une démarche de sélection de modèle, quelle valeur de K choisiriez-vous ?



3) *2pt* On s'intéresse à un problème de classification dans lequel des observations $\mathbf{x} = (x_1, x_2)$ de dimension 2 appartiennent à deux classes $\mathscr{C}_1$ et $\mathscr{C}_2$.

On considère le classifieur naïf de Bayes *uniforme*, dans lequel les probabilités conditionnelles suivent les lois uniformes suivantes :

$$p(x_1|\mathscr{C}_1) = \begin{cases} \frac{1}{b_1^1 - a_1^1} & \text{si } x \in [a_1^1, b_1^1] \\ 0 & \text{sinon} \end{cases}$$

$$p(x_2|\mathscr{C}_1) = \begin{cases} \frac{1}{b_2^1 - a_2^1} & \text{si } x \in [a_2^1, b_2^1] \\ 0 & \text{sinon} \end{cases}$$

$$p(x_1|\mathscr{C}_2) = \begin{cases} \frac{1}{b_1^2 - a_1^2} & \text{si } x \in [a_1^2, b_1^2] \\ 0 & \text{sinon} \end{cases}$$

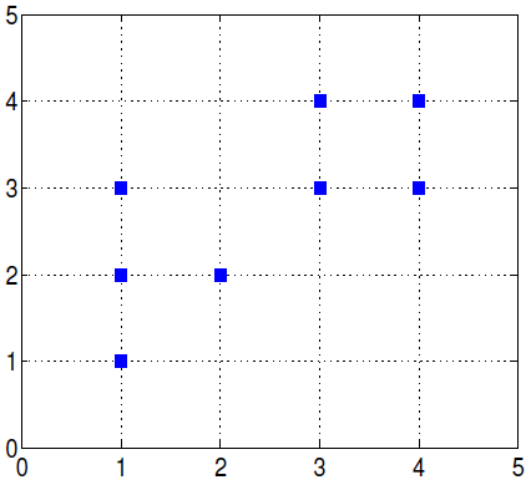
$$p(x_2|\mathscr{C}_2) = \begin{cases} \frac{1}{b_2^2 - a_2^2} & \text{si } x \in [a_2^2, b_2^2] \\ 0 & \text{sinon} \end{cases}$$

Bien entendu, on suppose $a_1^1 < b_1^1$, $a_2^1 < b_2^1$, $a_1^2 < b_1^2$, et $a_2^2 < b_2^2$.

Quelle est la classe prédite pour une nouvelle observation $\mathbf{x}$? Vous ferez un dessin et serez amené à discuter la position relative des deux rectangles $R^1 = [a_1^1, b_1^1] \times [a_2^1, b_2^1]$ et $R^2 = [a_1^2, b_1^2] \times [a_2^2, b_2^2]$.
Indication : la prédiction dépend uniquement de l'appartenance de $\mathbf{x}$ à R^1 et/ou R^2.

Comment se simplifie la règle de décision lorsque les probabilités a priori sont égales?

4) *1pt* Cette question porte sur un problème de partitionnement. On considère des observations appartenant à un espace de dimension 2, représentées par des points sur le graphique suivant (première dimension : axe horizontal; seconde dimension : axe vertical).



On veut identifier deux classes parmi ces observations en utilisant l'algorithme K-means avec $K = 2$. Les deux centres initiaux ont pour coordonnées $(2,3)$ et $(4,2)$. Représentez graphiquement chaque itération de l'algorithme. Quelle est la partition obtenue?

5) *2.5pt* On considère des points du plan séparés en deux classes : la classe 1 est formée des points de coordonnée $a = (-0,4;0,2)$, $b = (-0,5;0,3)$, $c = (0,8;0,2)$, $d = (1;0,6)$, et la classe 2 est formée des points de coordonnées $e = (-0,1;0,2)$, $f = (0,1;0,8)$, $g = (0,1;0,2)$.

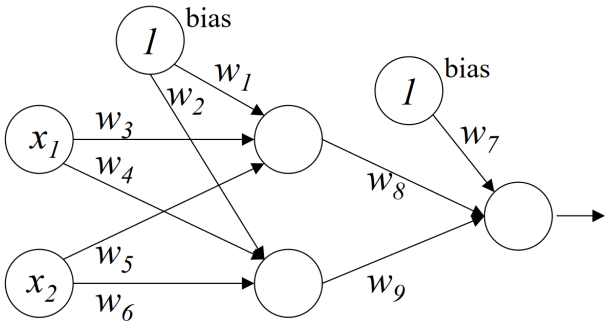
(a) Le classifieur de la régression logistique « standard » permet-il de séparer ces classes ?

(b) On considère le noyau k défini pour tout couple de points du plan $(x, x') \neq (0,0)$ par l'équation $k(x,x') = \frac{x \cdot x'}{||x||||x'||}$. Ici $x \cdot x'$ désigne le produit scalaire euclidien (classique), et $||x||$ la norme euclidienne. Trouvez une fonction ϕ de $\mathbb{R}^2$ dans $\mathbb{R}^2$ telle que $k(x,x') = \phi(x) \cdot \phi(x')$. Représentez les observations dans l'espace de redescription : vous représenterez sur un même graphique dans $\mathbb{R}^2$ les observations $\mathbf{x}$ et leur image $\phi(\mathbf{x})$ dans l'espace de redescription.

(c) Constatez que les deux classes sont linéairement séparables dans l'espace de redescription, et tracez la droite de séparation entre les deux classes de marge maximale. (on se contentera d'une justification graphique, aucun calcul n'est demandé)
Quels sont les vecteurs supports ?

(d) Quelle est la frontière de séparation entre les deux classes dans l'espace des données d'origine induites par ce séparateur à vastes marges (modèle de SVM avec noyau k, pas de variables d'écart) ?

6) *1,5pt* On considère le réseau de neurones à une couche cachée de la figure suivante, pour un problème de classification de données $\mathbf{x} \in \mathbb{R}^2$ en deux classes C_1 et C_2.



La fonction d'activation des neurones de la couche cachée est linéaire : $\sigma(z) = cz$, avec c un paramètre positif. De manière à pouvoir interpréter la sortie du réseau comme la probabilité $p(C_1|\mathbf{x})$, le neurone de sortie a une fonction d'activation sigmoïde $\sigma_s(z) = \frac{1}{1+e^{-z}}$.

(a) Exprimez $p(C_1|\mathbf{x})$ en fonction des poids w_i, de c, et des composantes (x_1, x_2) de $\mathbf{x} \in \mathbb{R}^2$.

(b) Déterminez l'équation de la frontière de classification dans $\mathbb{R}^2$ de ce réseau. Quelle est sa nature géométrique ?

(c) Dessinez un réseau de neurones sans couche cachée qui fournit une frontière de séparation de même nature géométrique que celle de ce réseau.

Partie 2

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

Déposez votre carnet, avec ses cellules exécutées, au format HTML :

— sous Jupyter notebook : "`File`" → "`Download as`" → "`HTML`" ;

— sous VSCode : `Export` dans la barre juste au dessus du carnet (peut être caché dans "..." à dérouler), et choisir 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.