

## EXAMEN

Instructions : – Une feuille aide-mémoire recto verso manuscrite est permise ;  
 – Durée de l'examen : 2 h 50.

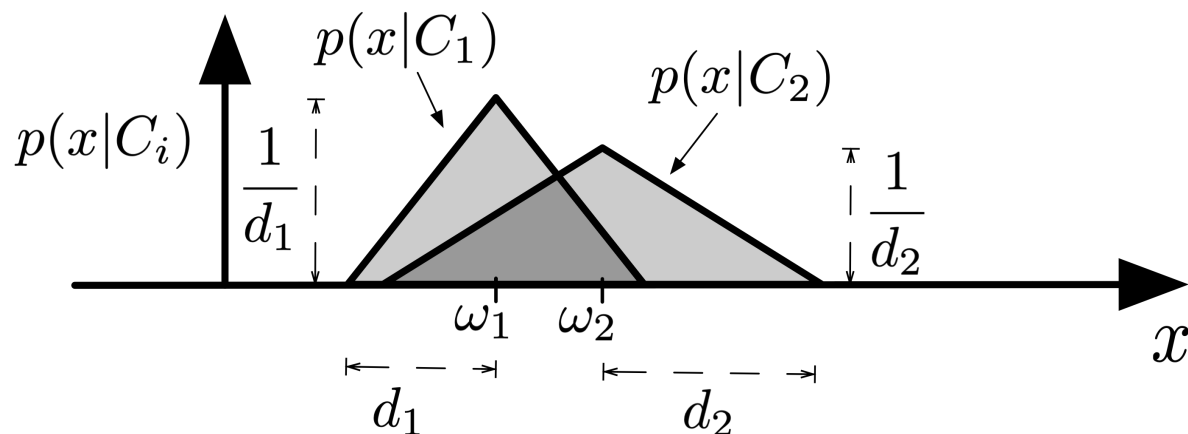
Pondération : Cet examen compte pour 40% de la note finale.

### Question 1 (20 points sur 100)

Soit un problème de classement à deux classes et en une dimension, dont les vraisemblances de classe sont décrites par les densités de probabilité suivantes, paramétrée chacune par un centre  $\omega_1$  et une demi-largeur  $d_i$  :

$$p(x|C_1) = \begin{cases} \frac{x+\omega_1-d_1}{d_1^2} & x \in [\omega_1 - d_1, \omega_1[ \\ \frac{-x+\omega_1+d_1}{d_1^2} & x \in [\omega_1, \omega_1 + d_1] \\ 0 & \text{autrement} \end{cases}, \quad p(x|C_2) = \begin{cases} \frac{x+\omega_2-d_2}{d_2^2} & x \in [\omega_2 - d_2, \omega_2[ \\ \frac{-x+\omega_2+d_2}{d_2^2} & x \in [\omega_2, \omega_2 + d_2] \\ 0 & \text{autrement} \end{cases}$$

qui sont représentées dans ce qui suit.

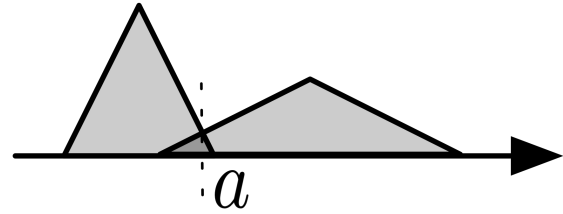


Sans perte de généralité, on suppose que  $\omega_2 > \omega_1$ .

- (9) (a) Supposons que les probabilités a priori sont égales ( $P(C_1) = P(C_2) = 0,5$ ). La règle de décision à utiliser pour assigner une donnée  $x$  à la classe  $C_1$  ou  $C_2$  selon cette modélisation change selon les configurations des densités, c'est-à-dire les valeurs de  $\omega_1$ ,  $\omega_2$ ,  $d_1$  et  $d_2$ . Voici une première configuration possible.

**Configuration 1 :**  $(\omega_1 - d_1) < (\omega_2 - d_2) < (\omega_1 + d_1) < (\omega_2 + d_2)$

$$h(x) = \begin{cases} C_1 & (\omega_1 - d_1) < x < a, \\ C_2 & a < x < (\omega_2 + d_2), \\ \text{indéfini} & \text{autrement.} \end{cases}$$



De façon similaire, donnez la règle de décision et le tracé des densités pour les trois autres configurations possibles.

- (5) (b) Pour la configuration 1 présentée précédemment avec probabilités a priori égales, mais où on suppose que  $\omega_1 < (\omega_2 - d_2) < (\omega_1 + d_1) < \omega_2$ . Déterminez l'équation pour calculer la valeur de  $a$  qui devrait être utilisée comme seuil dans la règle de décision.
- (6) (c) Toujours selon la configuration 1 et  $\omega_1 < (\omega_2 - d_2) < (\omega_1 + d_1) < \omega_2$ , déterminez l'équation pour calculer le taux d'erreur bayésien optimal avec cette modélisation, c'est-à-dire le taux d'erreur que l'on obtiendrait pour classer des données suivant ces densités.

## Question 2 (20 points sur 100)

Soit le jeu de données suivant  $\mathcal{X} = \{(\mathbf{x}^t, r^t)\}_{t=1}^8$ , en deux dimensions :

$$\begin{aligned} \mathbf{x}^1 &= [0,65 \ 0,6]^\top, & \mathbf{x}^2 &= [0,25 \ 0,55]^\top, & \mathbf{x}^3 &= [0,2 \ 0,7]^\top, & \mathbf{x}^4 &= [0,4 \ 0,5]^\top, \\ \mathbf{x}^5 &= [0,7 \ 0,35]^\top, & \mathbf{x}^6 &= [0,4 \ 0,3]^\top, & \mathbf{x}^7 &= [0,15 \ 0,35]^\top, & \mathbf{x}^8 &= [0,1 \ 0,45]^\top. \end{aligned}$$

Les étiquettes de ces données sont  $r^1 = r^2 = r^3 = r^4 = -1$  et  $r^5 = r^6 = r^7 = r^8 = 1$ .

La matrice des distances  $D_{\mathcal{X}}$  donne la distance euclidienne entre toutes les paires possibles de données :

$$D_{\mathcal{X}} = \begin{bmatrix} d(\mathbf{x}^1, \mathbf{x}^1) & d(\mathbf{x}^1, \mathbf{x}^2) & \cdots & d(\mathbf{x}^1, \mathbf{x}^N) \\ d(\mathbf{x}^2, \mathbf{x}^1) & d(\mathbf{x}^2, \mathbf{x}^2) & \cdots & d(\mathbf{x}^2, \mathbf{x}^N) \\ \vdots & \vdots & \ddots & \vdots \\ d(\mathbf{x}^N, \mathbf{x}^1) & d(\mathbf{x}^N, \mathbf{x}^2) & \cdots & d(\mathbf{x}^N, \mathbf{x}^N) \end{bmatrix}$$

$$= \begin{bmatrix} 0,0000 & 0,4031 & 0,4610 & 0,2550 & 0,2550 & 0,3905 & 0,5590 & 0,5701 \\ 0,4031 & 0,0000 & 0,1581 & 0,1500 & 0,4924 & 0,2915 & 0,2236 & 0,1803 \\ 0,4610 & 0,1581 & 0,0000 & 0,2500 & 0,6103 & 0,4472 & 0,3536 & 0,2693 \\ 0,2550 & 0,1500 & 0,2500 & 0,0000 & 0,3606 & 0,2500 & 0,3202 & 0,3162 \\ 0,2550 & 0,4924 & 0,6103 & 0,3606 & 0,0000 & 0,3041 & 0,5500 & 0,6083 \\ 0,3905 & 0,2915 & 0,4472 & 0,2500 & 0,3041 & 0,0000 & 0,2550 & 0,3354 \\ 0,5590 & 0,2236 & 0,3536 & 0,3202 & 0,5500 & 0,2550 & 0,0000 & 0,1118 \\ 0,5701 & 0,1803 & 0,2693 & 0,3162 & 0,6083 & 0,3354 & 0,1118 & 0,0000 \end{bmatrix}.$$

- (8) (a) Tracez les régions de décision selon les données d'entraînement présentée en préambule de la question pour un classifieur de type plus proche voisin (avec un seul voisin,  $k = 1$ ) utilisant une distance euclidienne. Tracez le tout dans la **feuille de réponse fournie** à la fin du questionnaire. Donnez également le taux de classement selon une méthodologie *leave-one-out* avec cette configuration, sur ces données. Utilisez l'indice de donnée de valeur la plus faible pour briser les cas d'égalité de distances.

- (12) (b) Nous obtenons le résultat suivant en effectuant l'entraînement d'un SVM linéaire à **marge douce** avec ces données, avec comme valeur de paramètre de régularisation  $C = 100$  :

$$\begin{aligned}\alpha^1 &= 0,00, & \alpha^2 &= 51,56, & \alpha^3 &= 0,00, & \alpha^4 &= 100,00, \\ \alpha^5 &= 58,44, & \alpha^6 &= 0,00, & \alpha^7 &= 0,00, & \alpha^8 &= 93,11, \\ w_0 &= 8,47\end{aligned}$$

En utilisant la **feuille de réponse fournie** à la fin du questionnaire, tracez les éléments suivants :

- Frontière de décision du SVM (droite continue, —);
- Frontières de la marge (droites en traits pointillés, ---);
- Vecteurs de support (encerclez les points, ○);
- Données dans la marge (encadrez les points, □).

### Question 3 (20 points sur 100)

Supposons que l'on veut entraîner un autoencodeur à deux couches, la première couche comme encodeur et l'autre comme décodeur, toutes deux déterminée selon un apprentissage en ligne par rétropropagation des erreurs. Les poids de ces deux couches sont distincts pour les besoins de l'entraînement.

Dans ce réseau, une fonction de transfert linéaire est utilisée ( $f_{\text{lin}}(x) = x$ ). La sortie d'un neurone de la couche d'encodage se modélise comme suit :

$$z_i^t = (\mathbf{w}_i^{\text{enc}})^\top \mathbf{x}^t + w_{i,0}^{\text{enc}}, \quad i = 1, \dots, K.$$

La sortie de la couche de décodage se modélise comme suit, ce qui correspond à la donnée d'entrée reconstruite :

$$\hat{x}_j^t = (\mathbf{w}_j^{\text{dec}})^\top \mathbf{z}^t + w_{j,0}^{\text{dec}}, \quad j = 1, \dots, D.$$

Le critère de performance utilisé est l'erreur quadratique de reconstruction  $E_{\text{rec}}^t$ , calculée sur une donnée  $\mathbf{x}^t$  :

$$e_j^t = (x_j^t - \hat{x}_j^t), \quad E_{\text{rec}}^t = \frac{1}{2} \|\mathbf{x}^t - \hat{\mathbf{x}}^t\|^2 = \frac{1}{2} \sum_{j=1}^D (e_j^t)^2.$$

Répondez aux questions suivantes sur l'entraînement d'un tel réseau de neurones.

- (10) (a) Détaillez les équations pour mettre à jour les poids des neurones de la couche du décodeur, par descente du gradient et en utilisant l'erreur de reconstruction  $E_{\text{rec}}^t$  d'une donnée.
- (10) (b) Détaillez maintenant les équations pour mettre à jour les poids des neurones de la couche de l'encodeur, toujours par descente du gradient et en utilisant l'erreur de reconstruction  $E_{\text{rec}}^t$  d'une donnée.

**Question 4** (40 points sur 100)

Répondez aussi brièvement et clairement que possible aux questions suivantes.

- (4) (a) Voici la formalisation mathématique du théorème de Bayes pour le classement, tel que présenté en classe.

$$P(C_i|\mathbf{x}) = \frac{P(C_i) p(\mathbf{x}|C_i)}{p(\mathbf{x})}$$

Définissez textuellement ce que signifie chacun des termes mathématiques suivants :  $P(C_i)$ ,  $p(\mathbf{x}|C_i)$ ,  $p(\mathbf{x})$  et  $P(C_i|\mathbf{x})$ .

- (4) (b) Expliquez quel sera le niveau de performance et la complexité attendue d'un modèle d'apprentissage supervisé lorsque celui-ci est dit à biais faible et variance élevée.
- (4) (c) Dans un contexte de classement paramétrique avec modèles de lois normales multivariées, indiquez quelle(s) simplification(s) aux matrices de covariance permet d'obtenir un modèle de classement linéaire.
- (4) (d) Dans la méthode de classement par les  $k$ -plus proches voisins, expliquez l'effet de l'augmentation de la valeur de  $k$  (nombre de voisins) sur les frontières de décision et les performances, supposant que l'on a un  $N$  (taille du jeu d'entraînement) faible. Précisez également ce qui arrive dans le cas limite où la valeur de  $k$  se rapproche de celle de  $N$ .
- (4) (e) Expliquez pourquoi le critère des moindres carrés est préférable au critère du perceptron pour des modèles de discriminants linéaires devant être entraînés sur des données non linéairement séparables.
- (4) (f) Expliquez comment un SVM à noyau est en mesure de traiter avec succès des données de classement non linéairement séparables en comparaison avec un SVM linéaire.
- (4) (g) Expliquez pourquoi l'utilisation d'un jeu de données de validation, indépendant du jeu d'entraînement et de test, est nécessaire pour arrêter l'entraînement et sélectionner le modèle, avec des réseaux de neurones tels que les perceptrons multicouches.
- (4) (h) Il a été dit dans le cours que les réseaux de neurones profonds modernes sont des modèles qui possèdent des mécanismes leur permettant de bien gérer les données à haute dimensionnalité. Expliquez comment ceci est effectué concrètement dans les réseaux de neurones à convolution, par exemple pour traiter des images.
- (4) (i) Expliquez la façon la plus courante d'effectuer un transfert de représentation avec des réseaux de neurones profonds, dans un contexte de classement.
- (4) (j) Lorsqu'on implémente un modèle de classement supervisé avec scikit-learn, indiquez le nom des trois méthodes principales de l'interface que l'on doit définir, en décrivant la fonction de chacune d'elle en une phrase.