Partiel Machine Learning - DAC

2 heure - Barème indicatif - Document : 1 à 2 feuilles A4
Le barème est donné à titre indicatif pour indiquer l'importance relative des exercices.

Exercice 1 (7 points) – Vrai ou faux

Pour chaque question, une seule réponse est correcte. Donner vos réponses, **sans justifier**, sous la forme 1-V par exemple si l'affirmation 1 est vrai, 1-F si l'affirmation est fausse. Chaque bonne réponse apporte 0.25 points, chaque mauvaise retire 0.25 points à votre note (barème indicatif).

V 1. Le classifieur naïf bayésien est meilleur que le classifieur bayésien.

2. La méthode des histogrammes est moins coûteuse en temps de calcul pour l'inférence que les fenêtres de Parzen.

V 3. Il vaut mieux utiliser les fenêtres de Parzen en grande dimension que la méthode des histogrammes.

4. Un choix de longueur de discrétisation trop petit dans la méthode des histogrammes peut conduire à du sous apprentissage.

5. Il n'est pas possible de contrôler le sur-apprentissage dans la méthode des fenêtres de Parzen.

6. La densité d'une variable aléatoire est toujours inférieure à 1.

7. Un arbre de décision peut séparer exactement tout ensemble de points disjoints.

F 8. Il est nécessaire de normaliser les données lorsqu'on apprend un arbre de décision.

9. Le signe de l'entropie permet de déterminer l'aléa dans un ensemble de données.

V 10. La vraisemblance d'un jeu de données par rapport à un modèle idéal baisse lorsque le nombre de données augmente.

11. Selon l'algorithme de descente de gradient, il est toujours possible de trouver, si le gradient est non nul, un pas d'apprentissage ϵ permettant de faire décroître le coût à minimiser.

12. Lorsqu'on remarque qu'on sur-apprend en utilisant un K-NN, il vaut mieux augmenter K.

F 13. La régression logistique est adaptée pour prédire la note d'un film en en fonction de ses caractéristiques.

V 14. Lorsque les données sont de petites dimensions, le risque de sur-apprentrissage est plus grand.

F 15. La régression logistique est capable de traiter plus de problèmes que le perceptron.

16. Régulariser la régression linéaire : $argmin_{\mathbf{w},b}||\mathbf{w}^t X + b - Y||^2 + \lambda||\mathbf{w}||^2$ permet de garantir de n'avoir qu'une unique solution.

17. Il est possible d'utiliser une régression linéaire pour apprendre un problème de type $y_i = e^{\beta x_i + \epsilon_i}$ avec ϵ_i un bruit gaussien.

18. Il vaut mieux initialiser à 0 les poids d'un perceptron en début d'apprentissage pour des raisons de stabilité.

F 19. Les classifieurs linéaires $\mathbf{w}_1 = (1, -3, -2)$ et $\mathbf{w}_2 = (5, -15, 10)$ sont équivalents.

20. La minimisation de $\sum_{i|y^i=1} \lambda_+ \mathbf{1}_{f(\mathbf{x}^i)<0} + \sum_{i|y^i=-1} \lambda_- \mathbf{1}_{f(\mathbf{x}^i)\geq 0}$ selon le classifieur f avec $\mathbf{1}$ la fonction indicatrice et $\lambda_+ > \lambda_-$, favorise la classification des exemples comme positif plutôt que négatif.

21. La dimensionnalité de la projection d'un noyau polynomial est linéaire par rapport à d le degré du polynôme.

22. La complexité du calcul d'un noyau polynomial est linéaire par rapport à d le degré du polynôme.

23. Dans un noyau gaussien, un petit σ - la variance - permet de baisser le sur-apprentissage.

24. Pour un jeu de données linéairement séparable, on obtient la même frontière de décision si on entraîne un SVM sur toutes les données ou seulement sur les vecteurs supports trouvés lors de l'apprentissage sur toutes les données.

25. Lorsque les données sont linéairement séparables, un perceptron est équivalent à un SVM linéaire.

26. Certaines frontières apprises avec un perceptron ne peuvent être apprises par un SVM avec un noyau polynomial.

27. Lors d'une descente batch de gradient, il n'est pas nécessaire de mélanger les exemples.

28. Pour un réseau de neurones à au moins deux couches cachées, ajouter des neurones dans une couche ne permet pas d'augmenter les performances.
29. Un coût $max(0, 1 - f(x)y)$ est généralement meilleur que $max(0, 0.1 - f(x)y)$ car il permet d'augmenter la marge des données avec les frontière de décision.
30. Pour apprendre un réseau de neurones, il suffit de connaître le gradient de la fonction de coût par rapport aux entrées de chaque couche et par rapport aux paramètres de chaque couche.

Exercice 2 (4 points) – Régression logisique économe

On considère le problème de régression logistique linéaire (sans biais pour simplifier) pour la classification binaire entre deux classes $\{-1, +1\}$. On rappelle l'hypothèse dans ce contexte : $p(y = +1|\mathbf{x}) = \frac{1}{1+e^{-f_{\mathbf{w}}(\mathbf{x})}}$. On considère un ensemble de données $D = \{(\mathbf{x}^i, y^i)\}_{i=1}^N$ avec $\mathbf{x}^i \in \mathbb{R}^d$ et $y^i \in \{-1, +1\}$, avec $\mathbf{x}^i = (x_1^i, \ldots, x_d^i)$.

Q 2.1 Donner l'expression de la vraisemblance $L(D; \mathbf{w})$ sur le jeu de données D, puis celle du logarithme de la vraisemblance $LL(D; \mathbf{w})$ en fonction de $f_{\mathbf{w}}(\mathbf{x}^i)$ et de y^i. Cherche-t-on à la maximiser ou à la minimiser ?

Q 2.2 On veut pénaliser la log-vraisemblance $LL(D; \mathbf{w})$ du modèle par une pénalité $L2$ sur le vecteur de poids $\mathbf{w}$, soit par $\frac{\lambda}{2}\|\mathbf{w}\|^2$ avec λ la constante de pénalisation. Donner le coût à minimiser en fonction de y^j, x_i^j et des w_i les coordonnées de $\mathbf{w}$ (on rappelle que $w_0 = 0$ pour simplifier).
On considère dans la suite que les exemples $\mathbf{x}^i$ sont très parcimonieux : pour chaque exemple, il n'y a que s dimensions en moyenne non nulles, avec s très inférieure à d.

Q 2.3 Donner dans le cas $\lambda = 0$ la règle de mise-à-jour pour l'algorithme de descente de gradient pour un pas de gradient ϵ sur l'exemple $(\mathbf{x}^i, y^i)$ (descente stochastique). Quelle est la complexité computationnelle en fonction de s ?

Q 2.4 On considère $\lambda > 0$, donner dans ce cas la règle de mise-à-jour.

Q 2.5 Soit $\mathbf{w}^t$ le paramètre à l'itération t, on considère k itérations supplémentaire de descente de gradient stochastique jusqu'à $t + k$. Exprimer $\mathbf{w}_i^{t+k}$ en fonction de $\mathbf{w}_i^t$, k, ϵ et λ si on ne considère que des exemples tels que $\mathbf{x}_i^j = 0$ durant ces k itérations.

Q 2.6 Proposer un algorithme d'apprentissage efficace dans le cas d'exemples parcimonieux. Quel est le temps moyen par exemple ?

Exercice 3 (4 points) – Risque balancé

On considère un problème de classification à deux classes C_1 et C_2, on suppose $p(x|C_k) = \frac{2x}{a_k^2} e^{-\frac{x^2}{a_k^2}}$ avec $a_k \in \mathbb{R}$ les deux paramètres inconnus et $x \in \mathbb{R}^+ - \{0\}$. Soit $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^N$ un ensemble d'apprentissage. On notera N_k le nombre d'exemples de la classe C_k.

Q 3.1 Pour une classe C_k donnée, on veut déterminer à partir des données le paramètre a_k de sa loi par maximum de vraisemblance.

Q 3.1.1 Donner l'expression de la log-vraisemblance pour le paramètre a_k de la classe C_k.

Q 3.1.2 En déduire l'estimation de $\hat{a}_k$.

Q 3.2 On utilise un classifieur bayésien pour classer les données, mais avec des coûts asymétriques : $l_{12} = 2\beta$ si l'exemple était de classe C_2 mais qu'on a prédit C_1 et $l_{21} = \beta$ dans le cas contraire. On considère par ailleurs les classes équiprobables.

Q 3.2.1 Donner l'expression du risque $R(C_k|x)$ pour les deux classes.

Q 3.2.2 Quelle est la condition sur $P(C_1|x)$ et $P(C_2|x)$ pour classer x comme C_1 ?

Q 3.2.3 Donner l'expression de la frontière de décision en fonction de x et $\hat{a}_k$.

Exercice 4 (6 points) – **SVM ordonné**

On considère un problème d'ordonnancement : on considère une question et un ensemble de documents $\{\mathbf{x}^i\}_{i=1}^N$; certains documents sont plus pertinents que d'autres pour cette question. On note $\mathbf{x}^i \succ \mathbf{x}^j$ lorsque le document i est plus pertinent que le document j. On souhaite trouver une fonction $f : \mathbb{R}^d \to \mathbb{R}$ qui permette d'inférer l'ordonnancement à partir d'un document $\mathbf{x}$. En particulier, on souhaite que sur le jeu de données d'entraînement $f(\mathbf{x}^i) > f(\mathbf{x}^j)$ si $\mathbf{x}^i \succ \mathbf{x}^j$.
Pour simplifier le problème, on considère qu'en fait les documents sont partitionnés en deux ensembles : les documents pertinents $S_+ = \{\mathbf{u}^i\}_{i=1}^{n_+}$ et les documents non pertinents $S_- = \{\mathbf{v}^j\}_{j=1}^{n_-}$ avec $n_+ + n_- = n$. La comparaison de deux documents de S_+ ou de deux documents de S_- ne nous importent pas, par contre on souhaite que pour $\mathbf{u}^i \in S_+$ et $\mathbf{v}^j \in S_-$, $f(\mathbf{u}^i) - f(\mathbf{v}^j) > 0$. On considère dans la suite que f est linéaire et paramétrée par $\mathbf{w}$: $f_{\mathbf{w}}(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle$.
On définit le problème d'optimisation suivant :

$$\min_{\mathbf{w}, \xi_{ij}} \frac{1}{2}\|\mathbf{w}\|^2 + C \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \xi_{ij}$$

avec les contraintes $\langle \mathbf{w}, \mathbf{u}^i \rangle - \langle \mathbf{w}, \mathbf{v}^j \rangle \geq 1 - \xi_{ij}$ et $\xi_{ij} \geq 0$ $\forall i = 1, \ldots, n_+$, $\forall j = 1, \ldots, n_-$.

Rappel : Pour un problème d'optimisation $min_{\theta \in \mathbb{R}^d} J(\theta)$ sous les m contraintes $g_j(\theta) \leq 0$ pour $j = 1 \ldots m$, le Lagrangien associé est $\mathcal{L}(\theta, \mu) = J(\theta) + \sum_{j=1}^{m} \mu_j g_j(\theta)$ avec $\mu_j \geq 0$. Les conditions KKT spécifient qu'à l'optimum, nécessairement $\nabla \mathcal{L}(\theta, \mu) = 0$, $g_j(\theta) \leq 0$, $\mu_j \geq 0$, et $\mu_j g_j(\theta) = 0$ pour $j = 1 \ldots m$.

Q 4.1 En considérant chaque terme du problème d'optimisation, expliquez en quoi cette formulation répond bien au problème. À quoi sert C ?

Q 4.2 Donner le Lagrangien correspondant à ce problème.

Q 4.3 Écrire les conditions d'optimalité par rapport aux variables $\mathbf{w}$, $\mathbf{u}^i$ et $\mathbf{v}^j$ et en déduire l'expression de $\mathbf{w}$.

Q 4.4 Quel est le problème dual correspondant ?

Q 4.5 Parmi les situations suivantes, quelles sont celles qui correspondent à des paires de points $(\mathbf{u}^i, \mathbf{v}^j)$ supports ?

- $f(\mathbf{u}^i) - f(\mathbf{v}^j) = 1$
- $f(\mathbf{u}^i) - f(\mathbf{v}^j) > 1$
- $f(\mathbf{u}^i) - f(\mathbf{v}^j) = 0$
- $f(\mathbf{u}^i) - f(\mathbf{v}^j) < 1$

Q 4.6 On définit maintenant le couple $(\mathbf{z}^l, y^l)$ tel que $\mathbf{z}^l = \mathbf{u}^i - \mathbf{v}^j$ et $y^l = 1$ si $\mathbf{u}^i \succ \mathbf{v}^j$, 0 sinon.

Q 4.6.1 Montrez que le problème d'ordonnancement précédent équivaut à un problème classique SVM appliqué aux points $\mathbf{z}^l$.

Q 4.6.2 On suppose que $n_+ = n_- = 2n$, quelle est la complexité en termes de nombre de paramètres à apprendre du dual entre la formulation précédente (question 4.4) et celle-ci (le SVM classique selon les exemples $\mathbf{z}^l$) ?