

Examen Machine Learning - DAC

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

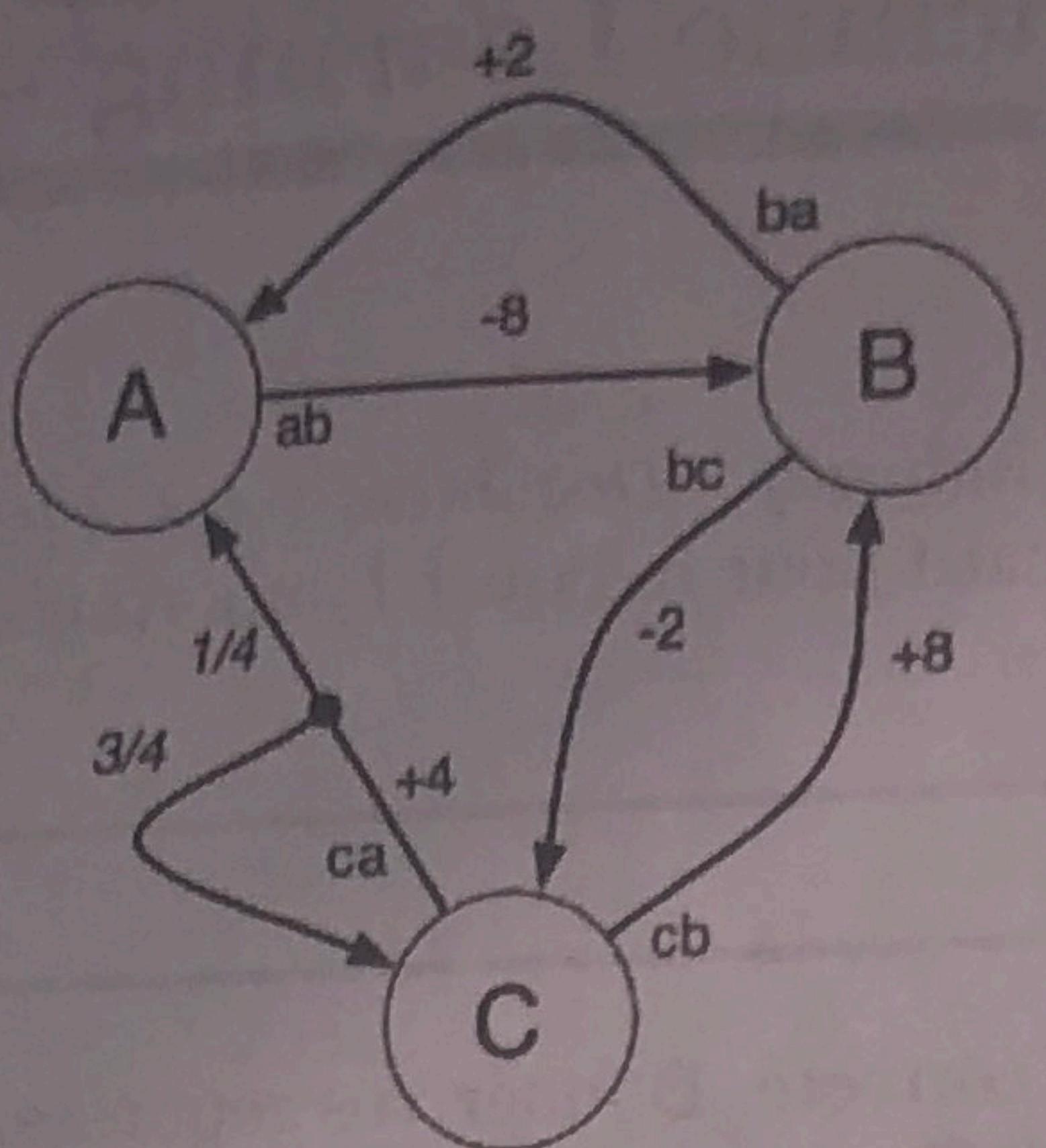
Exercice 1 (4 points) – Truth or Dare

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 (ou moins, barème indicatif).

1. Il est possible de multiplier deux noyaux admissibles pour former une nouvelle fonction noyau.
2. La régression logistique apprend des frontières non linéaires car la fonction logistique est non linéaire.
3. La fonction $k(x, x') = \max(0, x - x')$ est un noyau.
4. Le classifieur naïf bayésien peut classifier correctement le problème XOR (échiquier à 4 cases)
5. Entre deux modèles il est préférable de retenir celui qui a la plus petite erreur en apprentissage.
6. La somme de deux variables aléatoires gaussiennes suivent une loi gaussienne seulement si elles ont la même espérance.
7. La VC-dimension est utile pour déterminer le bruit dans les données
8. L'erreur en apprentissage diminue lorsque la VC-dimension de la famille de classificateurs considérée augmente.
9. Pour le boosting, à une itération donnée, tous les poids des exemples mal classés sont augmentés du même facteur multiplicatif.
10. Un SVM linéaire a une VC-dimension plus grande qu'un perceptron.
11. Une matrice de covariance diagonale implique l'indépendance des dimensions.
12. La distribution postérieure $P(\mu|X)$, selon des observations $X \sim \mathcal{N}(\mu; \sigma^2)$ avec σ^2 un paramètre de variance connu et $p(\mu)$ un prior gaussien, est toujours gaussienne.
13. Le fait qu'un prior soit conjugué à une vraisemblance indique que les deux distributions sont de la même famille.
14. L'incertitude d'une variable selon sa postérieure dépend de la variance de son prior.
15. Pour obtenir la distribution variationnelle optimale q_i^* d'un facteur i , il faut considérer l'espérance $E[\ln P(Z|X)]$ selon l'ensemble des facteurs de la distribution jointe Q .
16. En Reinforcement Learning, lorsque le MDP est parfaitement connu, il est possible de trouver la politique optimale sans exploration.

Exercice 2 (4 points) – Apprentissage par renforcement

On considère un MDP à 3 états représenté par le graphique ci-dessous : les lettres A, B, C dénotent les 3 états ; les arcs représentent les transitions possibles ; les couples de lettres minuscules ab, ba, bc, ca, cb représentent les actions possibles ; les entiers représentent les récompenses ; les fractions la probabilité de transition (seule l'action ca n'est pas déterministe).



Q 2.1 Rappeler la définition de la fonction valeur d'état $V^\pi(s)$ pour une politique π donné. A quoi sert la constante γ de cette fonction ?

Q 2.2 Donner l'équation de Bellman pour la fonction valeur d'état.

Q 2.3 Soit la politique π_1 telle qu'elle soit uniforme parmi les actions possibles à partir d'un état. En supposant toutes les valeurs initiales de V à 2, donner la première étape de l'évaluation de la politique π_1 (pour tous les états).

Q 2.4 En déduire une politique améliorée π_2 à partir des nouvelles valeurs de V .

Q 2.5 En supposant les valeurs initiales de V à 2, donner la première étape de l'algorithme value iteration. Est-ce que les nouvelles valeurs de V sont optimales ?

Exercice 3 (7 points) – Apprentissage par alternance

Dans cet exercice, on propose d'étudier deux algorithmes d'apprentissage de dictionnaire. Soit X une matrice de données de taille $N \times D$, où chaque ligne correspond à une donnée décrite sur D dimensions. Dans la suite, on notera pour une matrice M : $\mathbf{m}_{i,:}$ la i -ème ligne de M , $\mathbf{m}_{:,j}$ la j -ème ligne de M et $m_{i,j}$ un élément de M .

Q 3.1 L'objectif dans cette partie est de décomposer la matrice X en deux matrices $U \in \mathbb{R}^{N \times K}$ et $V \in \mathbb{R}^{D \times K}$ telles que $K << N$ et $X \approx UV^t$.

Soit le coût $C = C^1 + \lambda C^2$ avec $C^1 = \sum_{i=1}^N \|\mathbf{x}_{i,:} - \mathbf{u}_{i,:}V^t\|^2 = \sum_{i=1}^N C_i^1$ et $C^2 = \sum_{i=1}^N \|\mathbf{u}_{i,:}\|_1 = \sum_{i=1}^N C_i^2 = \sum_{i=1}^N \sum_{j=1}^K |u_{i,j}|$.

Q 3.1.1 Exprimer à l'aide des $u_{i,k}$ et $v_{:,k}$ le terme $\hat{\mathbf{x}}_i = \mathbf{u}_i V^t$.

Q 3.1.2 Décrire en quelques phrases l'objectif de l'apprentissage de dictionnaire. En particulier, à quoi correspondent les termes $u_{i,j}$ et $v_{:,k}$?

Q 3.1.3 Quelle est la signification du coût C_1 , du coût C_2 et du facteur λ ? Qu'obtient-t-on avec $\lambda = 0$? Et au contraire avec un λ très fort ?

Q 3.2 Optimisation du coût

Q 3.2.1 Donner l'expression de C^1 en fonction de $x_{i,j}$, $u_{i,k}$ et $v_{j,k}$.

Q 3.2.2 Le problème d'optimisation est-il convexe si on fixe U et que l'on cherche à optimiser que V ? Dans le cas inverse? Et si on cherche à optimiser simultanément U et V ?

Q 3.2.3 Calculer les dérivées $\frac{\partial C^1}{\partial u_{i,j}}$, $\frac{\partial C^1}{\partial v_{i,j}}$ et $\frac{\partial C^2}{\partial u_{i,j}}$.

Q 3.2.4 Donner un algorithme pour minimiser le coût $C(U, V)$. Converge-t-il vers un optimum global ou local ?

Exercice 4 (4 points) – La norme noyautée

On suppose les points suivants dans \mathbb{R}^2 : $\{(0.2, 0.4), (0.4, 0.8), (0.4, 0.2), (0.8, 0.4), (0, 0.4), (0.4, 0)\}$ tous de la classe 1, et $\{(0.4, 0.4), (0.8, 0.8)\}$ de la classe -1. On considère le noyau suivant : $k(\mathbf{x}, \mathbf{x}') = \frac{\langle \mathbf{x}, \mathbf{x}' \rangle}{\|\mathbf{x}\| \|\mathbf{x}'\|}$.

Q 4.1 A quelle fonction de projection $\phi(\mathbf{x})$ correspond ce noyau ?

Q 4.2 Dessiner les points dans le plan. Pour chaque point \mathbf{x}_i , représenter le point projeté $\phi(\mathbf{x}_i)$. Dessiner la séparatrice linéaire dans l'espace projeté. Indiquer les vecteurs supports.

Q 4.3 Dessiner la séparatrice correspondante dans l'espace initial et justifier votre réponse.

Q 4.4 Est-il possible d'apprendre cette séparatrice avec un perceptron ? Justifier.

Exercice 5 (7 points) – May I have your attention please

Une représentation (ou *embedding*) est un vecteur dans un espace euclidien qui permet de représenter en particulier des éléments discrets. On considère dans la suite un espace de représentation de dimension D pour représenter du texte. Pour cela, un vocabulaire est d'abord construit avec tous les tokens (mots) que l'on veut représenter. À chaque token est ensuite associé une représentation dans \mathbb{R}^D . Un vocabulaire de taille V peut être ainsi encodé par une matrice \mathbf{X} de taille $V \times D$ où la ligne i contient la représentation du token i . On notera \mathbf{x}_i cette représentation.

Pour simplifier, on considérera dans la suite que un token n'est présent qu'une seule fois dans un texte. On notera par abus de notation $|t|$ le nombre de tokens d'un texte et $\{i \in t\}$ les indexées des tokens dans le texte t .

Q 5.1 Une représentation naïve d'un texte t peut être obtenu en moyennant les représentations des mots du texte : la représentation \mathbf{z} du texte est ainsi $\mathbf{z} = \frac{1}{|t|} \sum_{i \in t} \mathbf{x}_i$. Pour réaliser la classification binaire, on utilise ensuite un perceptron (sans biais) avec un vecteur de poids $\mathbf{w} \in \mathbb{R}^D$ dans lequel on passe la représentation moyennée du texte suivi d'une sigmoïde et d'un coût cross-entropique.

Rappel : $\sigma(x) = \frac{1}{1+e^{-x}}$ et $CE(\hat{y}, y) = -(y \log(\hat{y}) + (1-y) \log(1-\hat{y}))$ avec $y \in \{0, 1\}$ et \hat{y} la prédiction.

Q 5.1.1 Donner l'expression de la sortie \hat{y} en fonction de \mathbf{z} la représentation moyenne du texte passée en entrée et l'expression du coût en fonction de \mathbf{z} sous la forme $y \log(1+\alpha) + (1-y) \log(1+\beta)$.

Q 5.1.2 Donner les expressions de $\nabla_{\mathbf{w}} CE(\mathbf{z}, y)$ et $\nabla_{\mathbf{z}} CE(\mathbf{z}, y)$. Puis donner l'expression de $\nabla_{\mathbf{x}_i} CE$ pour un mot i présent dans le texte t .

Q 5.1.3 On suppose qu'il n'y a que quelques mots qui sont importants dans le texte pour la classification. Quel(s) problème(s) pose ce modèle ?

Q 5.2 Pour pondérer les mots importants, on propose d'utiliser un modèle d'attention globale : connaissant le texte, une probabilité d'attention $\alpha_i = p(a_i | t)$ est portée sur chaque token du texte (il s'agit d'une distribution, leur somme est donc de 1). La représentation du texte sera alors $\mathbf{z} = \sum_{i \in t} \alpha_i \mathbf{x}_i$.

Pour modéliser cette attention, on utilise dans ce modèle simple une représentation globale de la tâche de classification sous la forme d'un vecteur $\mathbf{q} \in \mathbb{R}^D$ dans le même espace de représentation que les tokens et la pertinence d'un token \mathbf{x}_i avec la tâche est modélisée par le produit scalaire : $\mathbf{q} \cdot \mathbf{x}_i$. La représentation \mathbf{q} de la tâche doit être également apprise durant l'optimisation.

Q 5.2.1 Il faut tout d'abord transformer pour un texte la collection $\mathbf{q} \cdot \mathbf{x}_i$ en une distribution de probabilité $\{\alpha_i\}$. On souhaite avoir la relation suivante : $\log \alpha_i = K + \mathbf{q} \cdot \mathbf{x}_i$, avec K une constante pour le texte donné. Quelle fonction permet d'obtenir les α_i à partir des $\mathbf{q} \cdot \mathbf{x}_i$? (penser à passer à l'exponentiel et normaliser).

Q 5.2.2 Donner l'expression des gradients nécessaires à l'optimisation de ce problème.

Q 5.2.3 On souhaite que la représentation de la tâche \mathbf{q} ne soit pas globale pour toute la tâche mais dépende également du texte. Expliquer intuitivement l'intérêt de faire dépendre la représentation du texte considéré. Proposer une solution pour ce faire.