# 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.

## cice 1 (7.5 points) – Vrai ou faux

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

L'erreur d'apprentissage converge vers la vraie erreur lorsque le nombre de données en apprentissage tend vers l'infini si les données sont i.i.d.

Le nombre de sous-chaînes commune de tailles de 1 à $k$ entre deux séquences de caractères est un noyau admissible que l'on peut utiliser dans un SVM

Il est possible de multiplier deux noyaux admissibles pour former une nouvelle fonction noyau.

Plus l'ensemble d'apprentissage est grand, plus le risque de sur-apprentissage est petit.

La régression logistique apprend des frontières non linéaires car la fonction logistique est non linéaire.

La fonction $k(x, x') = max(0, x - x')$ est un noyau.

Le classifieur naïf bayésien peut classifier correctement le problème XOR (échiquier à 4 cases)

Il existe des noyaux qui ne sont pas des produits scalaires de projections en dimension finie

L'erreur en apprentissage de l'algorithme $k$-nn avec $k = 1$ est de 0

La rétro-propagation permet d'apprendre un optimum global d'un réseau de neurones

Entre deux modèles il est préférable de retenir celui qui a la plus petite erreur en apprentissage.

La somme de deux variables aléatoires gaussiennes suivent une loi gaussienne seulement si elles ont la même espérance.

La VC-dimension est utile pour déterminer le bruit dans les données

L'erreur d'apprentissage diminue lorsque la VC-dimension de la famille de classifieurs considérée augmente.

Pour le boosting, à une itération donnée, tous les poids des exemples mal classés sont augmentés du même facteur multiplicatif.

Un SVM linéaire a une VC-dimension plus grande qu'un perceptron.

Pour la régression linéaire, mettre un a priori gaussien sur les poids du modèle est équivalent à une régularisation L2.

Il n'est pas possible d'utiliser le kernel trick noyaux pour la régression linéaire.

On appelle bruit gaussien une perturbation des étiquettes qui suit une loi normale dont l'espérance dépend des données.

Une matrice de covariance diagonale implique l'indépendance des dimensions.

La distribution postérieure $P(\mu|X)$, selon des observations $X \sim \mathcal{N}(\mu; \sigma^2)$ avec $\sigma^2$ un paramètre de variance connu et $p(\mu)$ un prior gaussien, est toujours gaussienne.

Le fait qu'un prior soit conjugué à une vraisemblance indique que les deux distributions sont de la même famille.

Si deux distributions sont identiques, elles ont une entropie minimale.

Maximiser l'ELBO $L(Q)$ selon $Q$ permet de maximiser l'évidence P(X).

La divergence de Kullback-Leibler $D_K L(QZ||PZ|X)$ peut s'écrire : $-\int Q_Z \log \frac{PZ|X}{Q(Z)} dZ$.

L'espérance de la distribution postérieure prédictive $p(\tilde{x}|X)$, avec tout $x \sim \mathcal{N}(\mu, \sigma^2)$ (avec $\sigma^2$ connu), peut s'estimer en échantillonnant $N$ fois $\mu$ selon $p(\mu|X)$ et en considérant $\frac{1}{N} \sum_i \mathcal{N}(\tilde{x}; \mu^{(i)}, \sigma^2)$, avec $\mu^{(i)}$ le $i$-ème échantillon de $\mu$.

L'incertitude d'une variable selon sa postérieure dépend de la variance de son prior.

28. Pour obtenir la distribution variationnelle optimale $q_i^*$ d'un facteur $i$, il faut considérer l'espérance $E[lnP(Z|X)]$ selon l'ensemble des facteurs de la distribution jointe $Q$.

29. Une distribution $Q$ obtenue par inférence variationnelle tend à englober la distribution jointe qu'elle approxime.

30. L'approximation Mean-Field Relaxation correspond à décomposer en facteurs indépendants les variables d'une distribution jointe.

## Exercice 2 (4 points) – L'avis de Bayes

**Q 2.1** Soit un jeu de données $\{x^i, y^i\}_{i=1}^N$ avec $x^i$ et $y^i \in \mathbb{R}$.

**Q 2.1.1** Dans un premier temps, on utilise une méthode de régression linéaire pour résoudre le problème. Afin de tester la performance, on divise en deux le jeu de données, en un ensemble d'apprentissage et un ensemble de test. Comment vont se comporter les moyennes des erreurs en apprentissage et en test lorsque la taille de l'ensemble d'apprentissage augmente ? Justifiez.

**Q 2.1.2** On suppose maintenant que les données sont telles que $y^i \sim \mathcal{N}(log(wx^i), 1)$, le paramètre $w$ est inconnu, la variance est supposée de 1. Calculez la vraisemblance pour un paramètre $w$ donné sur le jeu de données. En déduire un algorithme pour l'estimation du paramètre.

**Q 2.2** On suppose $\{x^1, \ldots, x^N\}$ engendrés de manière i.i.d. de la distribution uniforme entre $-w$ et $w$ ($p(x) = 0$ si $|x| > w$, $p(x) = \frac{1}{2w}$ sinon). Donnez $w^*$ l'estimateur de maximum de vraisemblance de $w$.

**Q 2.3** On suppose les données suivantes décrites en 3 dimensions, et leur étiquette $y$ :

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x^1$ | 0 | 0 | 1 | 0 | 1 | 1 | 1 |
| $x^2$ | 0 | 1 | 1 | 0 | 1 | 0 | 1 |
| $x^3$ | 1 | 0 | 0 | 1 | 1 | 0 | 0 |
| $y$ | 0 | 0 | 0 | 1 | 1 | 1 | 1 |

Quelle serait la classification de l'exemple $(0, 0, 1)$ par un classifieur naïf bayésien ?

## Exercice 3 (4 points) – Le noyau c'est la norme

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 : $k(\mathbf{x}, \mathbf{x}') = \frac{<\mathbf{x}, \mathbf{x}'>}{\|\mathbf{x}\|\|\mathbf{x}'\|}$.

**Q 3.1** A quelle fonction de projection $\phi(\mathbf{x})$ correspond ce noyau ?

**Q 3.2** 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 3.3** Dessiner la séparatrice correspondante dans l'espace initial et justifier votre réponse.

**Q 3.4** Est-il possible d'apprendre cette séparatrice avec un perceptron ? Justifier.

## Exercice 4 (8 points) – Highway to gradient

L'architecture `Highway Network` a été proposé en 2015 spécifiquement pour les réseaux très profonds dédiés au traitement de l'image. Une couche de ce type de réseau est très semblable à celle d'un réseau classique mais réalise un mélange des entrées de la couche avec les sorties de cette couche.
Prenons par exemple une couche fully-connected non linéaire d'un réseau classique définie par la fonction $H(\mathbf{x}, \mathbf{W}_H) = \sigma(\mathbf{W}_H^t \mathbf{x} + b_H)$. Le Highway Network utilise une transformation $T(\mathbf{x}, \mathbf{W}_T) = \sigma(\mathbf{W}_T^t \mathbf{x} + b_T)$ afin de mélanger l'entrée $\mathbf{x}$ et la sortie usuelle de la couche $H(\mathbf{x}, \mathbf{W}_H)$ : la sortie $\mathbf{y}$ de la couche est (avec $\odot$ l'opérateur de produit terme à terme)

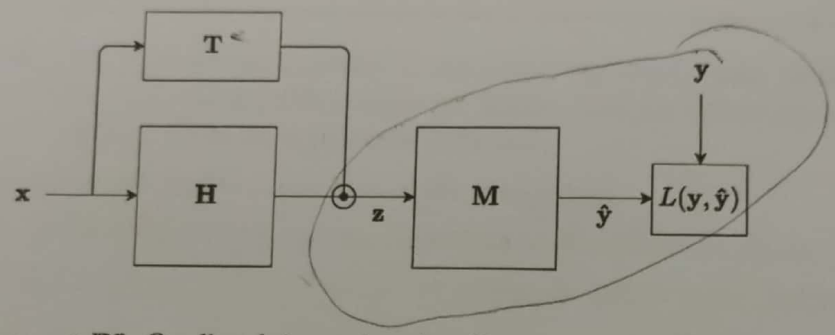
$$\mathbf{y} = H(\mathbf{x}, \mathbf{W}_H) \odot T(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \odot (1 - T(\mathbf{x}, \mathbf{W}_T))$$

**Q 4.1** Supposons que l'entrée $\mathbf{x}$ soit de dimension $d$ : $\mathbf{x} \in \mathbb{R}^d$. D'après la définition, quelles sont les dimensions de $\mathbf{W}_H$, $\mathbf{W}_T$, $\mathbf{y}$ ? (on supposera les biais $b_H$ et $b_T$ scalaires dans $\mathbb{R}$).

Q 4.2 Calculez la dérivée de la fonction sigmoïde. Dans la suite, vous pouvez la noter $\sigma'(x)$ sans développer. Rappel : $\sigma(x) = \frac{1}{1+e^{-x}}$

Q 4.3 On suppose le réseau de la figure ci-dessous, avec $\mathbf{z}$ la sortie d'une couche de Highway Network et $M$ une couche linéaire avec une fonction d'activation sigmoïde :

- $\mathbf{z} = H(\mathbf{x}, \mathbf{W}_H) \odot T(\mathbf{x}, \mathbf{W}_T) + \mathbf{x} \odot (1 - T(\mathbf{x}, \mathbf{W}_T))$
- $\hat{\mathbf{y}} = \sigma(\mathbf{W}_M^t \mathbf{z} + b_M)$
- $L(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{2}\|\mathbf{y} - \hat{\mathbf{y}}\|^2$




Q 4.3.1 On suppose que $\mathbf{y} \in \mathbb{R}^p$. Quelles doivent être les dimensions de $\mathbf{W}_M$ et $\hat{\mathbf{y}}$ ? (on suppose que le biais est scalaire dans $\mathbb{R}$).

Q 4.3.2 Calculez $\frac{\partial L(\mathbf{y},\hat{\mathbf{y}})}{\partial \hat{\mathbf{y}}_i}$ la dérivée du coût par rapport à la $i$-ème sortie du réseau.

Q 4.3.3 Calculez $\frac{\partial L(\mathbf{y},\hat{\mathbf{y}})}{\partial w_{i,j}^M}$ pour un poids $w_{i,j}^M$ de $\mathbf{W}_M$.

Q 4.3.4 Calculez $\delta_i = \frac{\partial L(\mathbf{y},\hat{\mathbf{y}})}{\partial z_i}$ pour la $i$-ème sortie $\mathbf{z}$ de la couche Highway.

Q 4.3.5 Calculez pour un poids $w_{i,j}^T$ $\frac{\partial L(\mathbf{y},\hat{\mathbf{y}})}{\partial w_{i,j}^T}$

Q 4.3.6 Calculez pour un poids $w_{i,j}^H$ $\frac{\partial L(\mathbf{y},\hat{\mathbf{y}})}{\partial w_{i,j}^H}$

Q 4.3.7 Donnez l'algorithme d'optimisation du réseau.

Q 4 (bonus) À votre avis, quel(s) problème(s) permet de résoudre un réseau Highway et pourquoi ?