MAPSI – Examen réparti 2 – 08/01/25 – 60 pts

Durée : 2h

Seuls documents autorisés : Calculatrice, 2 antisèches recto-verso
– Barème indicatif –

Exercice 1 (25 pts) – Régression linéaire sur-paramétrée

On considère un problème de régression multidimensionnel en imagerie médicale.
L'entrée du problème est une image médicale volumique, *e.g.*, un scanner, avec $\mathbf{x} \in \mathbb{R}^d$, d correspondant au nombre de voxels du volume, *e.g.*, $d = 512 \times 512 \times 100$. On cherche à prédire la densité des calcifications $y \in \mathbb{R}$.
On dispose d'un ensemble de N exemples d'entraînement $\{(\mathbf{x}_i, y_i)\}_{i\in[1..N]}$ pour estimer les paramètres du modèle. On va considérer un modèle de régression linéaire sur-paramétré, *i.e.*, pour lequel $d >> N$.

Modèle de prédiction

Q 1.1 Écrire la sortie prédite par le modèle de régression linéaire $\hat{y}_i$ par rapport à l'entrée $\mathbf{x}_i$ et le vecteur de paramètres $\mathbf{w}$, dont on précisera la dimension.

Q 1.2 En considérant l'ensemble des N exemples d'entraînement, écrire le vecteur de prédiction $\hat{\boldsymbol{Y}}$ en fonction des N entrées regroupées dans la matrice $\boldsymbol{X} \in \mathbb{R}^{N\times d}$.

Maximum de vraisemblance (MV). Avec ce modèle linéaire, on suppose que la probabilité de la sortie suit une loi normale dont la moyenne est donnée par la prédiction du modèle et de variance σ^2 : $P(y|\mathbf{x}, \mathbf{w}) \sim \mathcal{N}(\mathbf{x}^T\mathbf{w}, \sigma^2)$. On souhaite estimer les paramètres du modèle par maximisation de vraisemblance.

Q 1.3 En supposant que les N exemples d'entraînement soient indépendants et identiquement distribués (*i.i.d*), montrer que la log-vraisemblance des données s'écrit :

$$\mathcal{L}(\mathbf{w}) = -\frac{N}{2}log(2\pi\sigma^2) - \sum_{i=1}^{N}\frac{(y_i - \mathbf{x}_i^T\mathbf{w})^2}{2\sigma^2} \tag{1}$$

Q 1.4 On suppose que σ dans l'équation (1) est fixé et on s'intéresse à la maximisation de la log-vraisemblance par rapport à $\mathbf{w}$. Montrer que le problème est équivalent à minimiser la fonction de coût des moindres carrés, *i.e.*,

$$\mathbf{w}_{MV} = \arg\min_{\mathbf{w}} \mathcal{L}(\mathbf{w}) = \arg\min_{\mathbf{w}} ||Y - Xw||^2.$$

Q 1.4.1 Montrer que la fonction de coût précédente peut se réécrire sous la forme : $\mathcal{L}(\mathbf{w}) = \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{Y}$.

Q 1.4.2 En déduire que le gradient de la fonction de coût s'écrit : $\frac{\partial\mathcal{L}}{\partial\mathbf{w}} = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{Y}$.

Q 1.4.3 Avec $N << d$, existe-t-il une solution unique à l'annulation du vecteur gradient ?

Q 1.5 On cherche maintenant à optimiser l'équation (1) par rapport à σ.

Q 1.5.1 Écrire le gradient $\frac{\partial\mathcal{L}}{\partial\sigma}$

Q 1.5.2 Existe-t-il une solution analytique à l'annulation de ce gradient ? Si oui, l'expliciter et l'interpréter. Sinon, justifier.

Q 1.6 Ridge regression. On va ajouter une loi a priori gaussienne sur $\mathbf{w}$, *i.e.*, $p(\mathbf{w}) \sim \mathcal{N}(0, \sigma_p^2 I_d)$.

Q 1.6.1 Écrire la formulation du maximum a posteriori (MAP), *i.e.*, la log-vraisemblance $log[p(\mathbf{w}|(\mathbf{X}, \mathbf{Y}))]$

Q 1.6.2 On s'intéresse à l'optimisation du MAP par rapport à $\mathbf{w}$. Montrer que cette optimisation peut se réécriree comme la minimisation de la fonction de coût suivante :

$$\mathcal{L}_{ridge}(\mathbf{w}) = ||\mathbf{Y} - \mathbf{X}\mathbf{w}||^2 + \lambda||\mathbf{w}||^2. \tag{2}$$

Exprimer λ en fonction des données du problème.

Q 1.6.3 En déduire l'expression du vecteur gradient pour la régression linéaire *ridge*.

Q 1.6.4 Avec $N << d$, existe-t-il une solution unique à l'annulation du vecteur gradient ?

Q 1.6.5 Écrire la fonction python `ridge_regression(X,Y,lambda)` calculant le résultat de la régression *ridge*. Discuter la complexité de l'algorithme. Quelle alternative pourrait-on envisager ?

Exercice 2 (15 pts) – Apprentissage d'une Chaîne de Markov Cachée

On souhaite apprendre les paramètres (Π, A, B) d'une chaîne de Markov cachée (CMC) avec N états et M observations discrètes.

Q 2.1 Quel est le nombre de paramètres du modèle à estimer ?

Q 2.2 Vraisemblance d'une base de données de séquences observées
Pour apprendre les paramètres de la CMC, on dispose de K séquences observées, *i.e.*, $\mathbf{X}, \mathbf{S} := \{(\mathbf{x}^k, \mathbf{s}^k)\}_{k\in[1..K]}$ où chaque $(\mathbf{x}^k, \mathbf{s}^k)$ forme une séquence de T observations $\mathbf{x}^k := (x_1^k, ..., x_T^k)$ et de T états associés $\mathbf{s}^k := (s_1^k, ..., s_T^k)$. On suppose que les différentes séquences sont indépendantes et identiquement distribuées (*i.i.d*).

> **Rappel :** Vraisemblance d'une séquence observée
> On considère une séquence de T observations $\mathbf{x} := (x_1, ..., x_T)$ et d'états associés $\mathbf{s} := (s_1, ..., s_T)$. La probabilité jointe de $(\mathbf{x}, \mathbf{s})$ s'écrit alors :
>
> $$P(\mathbf{x}, \mathbf{s}|A, B, \Pi) = P(s_1|\Pi)\prod_{t=1}^{T}P(x_t|s_t, B)\prod_{t=2}^{T}P(s_t|s_{t-1}, A) \tag{3}$$

Q 2.2.1 Montrer que l'écriture de l'équation (3) se généralise ainsi :

$$P(\mathbf{X}, \mathbf{S}|A, B, \Pi) = \prod_{k=1}^{K}P(s_1^k|\Pi)\prod_{t=1}^{T}P(x_t^k|s_t^k, B)\prod_{t=2}^{T}P(s_t^k|s_{t-1}^k, A) = \prod_{k=1}^{K}\Pi_{s_1^k}\prod_{t=1}^{T}b_{s_t^k, x_t^k}\prod_{t=2}^{T}a_{s_{t-1}^k, s_t^k} \tag{4}$$

Q 2.2.2 Regrouper les termes des produits dans l'équation (4) et montrer qu'elle se réécrit :

$$P(\mathbf{X}, \mathbf{S}|A, B, \Pi) = \prod_{i=1}^{N} \Pi_i \prod_{i=1}^{M} b_{mi}^{h_{im}} \prod_{j=1}^{N} a_{ij}^{n_{ij}} \quad (5)$$

Que représentent l_i, h_{im} et n_{ij} ?

Q 2.2.3 En déduire la log-vraisemblance de l'équation (4).

Q 2.3 *Lagrangien*
Pour maximiser la log-vraisemblance par rapport à (A, B, Π), il faut ajouter les contraintes suivantes sur les paramètres à apprendre :

$$\sum_{i=1}^{N} \Pi_i = 1$$
$$\forall i, \sum_{m=1}^{M} b_{im} = 1$$
$$\forall i, \sum_{j=1}^{N} a_{ij} = 1$$

Nous allons utiliser le formalisme du Lagrangien pour cela.

Rappel : Pour résoudre $\text{argmax}_x\, g(x)$ sous la contrainte f(x) = b, on introduit le Lagrangien, qui s'écrit : $\mathcal{L}(x, \lambda) = g(x) - \lambda(f(x) - b)$, où λ est le multiplicateur de Lagrange. Une condition nécessaire d'optimalité revient à chercher la solution à $\frac{\partial \mathcal{L}}{\partial x} = 0$ et $\frac{\partial \mathcal{L}}{\partial \lambda} = 0$.

Q 2.3.1 Écrire le Lagrangien associé à l'apprentissage de la CMC.

Q 2.3.2 Résoudre le Lagrangien et interpréter le résultat de l'estimation des paramètres par maximum de vraisemblance.

Q 2.4 Écrire la fonction `CMC_MV(X, S, N, M)` qui réalise l'estimation des paramètres de la CMC par maximum de vraisemblance, à partir d'une liste de K séquences d'observation X et d'états associés S.

Q 2.5 Que se passe-t-il avec l'estimation précédente pour les transitions et émissions non observées ? Interpréter le problème et proposer une solution pour le mitiger.

Q 2.6 On dispose d'une nouvelle séquence de test $(r_1, \ldots, r_T)$ dont on cherche a prédire la séquence d'états cachés. Est-il possible de réaliser cela avec la CMC ? Si oui, préciser. Sinon, justifier. Quel est l'intérêt de la modélisation de la CMC et de l'entraînement réalisé ?

Exercice 3 (20pts+10pts) – Schéma de Monte-Carlo sur une base de données

Dans cet exercice, on ne considère que des variables aléatoires **discrètes**.

Soit une base de données $\mathcal{X}^{(N)} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(l)}, \ldots, \mathbf{x}^{(N)}\}$. Chaque $\mathbf{x}^{(l)} = (x_1^{(l)}, \ldots, x_j^{(l)}, \ldots, x_d^{(l)})$ est un vecteur de dimension d. On nomme donc $x_j^{(l)}$ la valeur se trouvant à la ligne l et la colonne j dans la base $\mathcal{X}^{(N)}$.
On suppose que les $\mathbf{x}^{(l)}$ sont indépendants, identiquement distribués (*i.i.d*) et suivent donc tous une distribution $\mathbb{P}(\mathbf{X})$ avec $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_d)$ variable aléatoire discrète de dimension d.
Enfin, on notera $\mathbf{x} = (x_1, \ldots, x_d)$ une valeur quelconque dans le domaine $D = D_1 \times \ldots \times D_d$ de $\mathbf{X}$.
De sorte qu'on peut écrire (en notant $\mathbb{P}(\mathbf{x}) := \mathbb{P}(\mathbf{X} = \mathbf{x})$) :

$$\sum_{\mathbf{x} \in D} \mathbb{P}(\mathbf{x}) = 1 \quad (7)$$

Dans un cas général, d peut être très grand et donc la somme de l'équation (7) peut être impossible à calculer en temps raisonnable.

Q 3.1 Schéma de Monte-Carlo
Soit une fonction $f : D \to \mathbb{R}$ et Y la variable aléatoire définie par $Y := f(X)$. On s'intéresse à la valeur $\mu_Y = \mathbb{E}_{\mathbb{P}}(f)$.

Q 3.1.1 En construisant un échantillon $\mathcal{Y}^{(N)} := \{y^{(l)} = f(x^{(l)}), \forall l \in \{1, \ldots, N\}\}$, montrer que le meilleur estimateur $\hat{\mu}_Y^{(N)}$ de μ_Y sur cet échantillon est :

$$\hat{\mu}_Y^{(N)} = \frac{1}{N} \sum_l y^{(l)}$$

Quel est le théorème qui nous permet de considérer que cet estimateur est sans biais ? i.e.

$$\lim_{N \to \infty} \hat{\mu}_Y^{(N)} = \mathbb{E}_{\mathbb{P}}(f) \quad (8)$$

Les conditions de ce théorème sont-elles satisfaites dans l'échantillon $\mathcal{Y}^{(N)}$?

Q 3.1.2 Écrire $\hat{\mu}_Y^{(N)}$ en fonction de $\mathcal{X}^{(N)}$. Que se passe-t-il quand $N \to \infty$? On appelle cette méthode d'approximation un schéma de Monte-Carlo. Proposer brièvement un algorithme (très simple) pour calculer cette valeur à partir de $\mathcal{X}^{(N)}$.

Q 3.2 Approximation de la distribution jointe $\mathbb{P}$
Pour tout $x \in D$, on note $n_{\mathbf{x}}$ le nombre d'occurrences de $\mathbf{x}$ dans la base $\mathcal{X}^{(N)}$. On note $\mathbb{P}^{(N)}(\mathbf{x})$ la distribution jointe empirique obtenue par les fréquences dans $\mathcal{X}^{(N)}$:

$$\forall \mathbf{x} \in D, \mathbb{P}^{(N)}(\mathbf{x}) = \frac{n_{\mathbf{x}}}{N}$$

Exemple : Avec $d = 2$, $\mathbb{P}$ uniforme sur X_1, X_2 binaires, $N = 5$ et la base de données $\mathcal{X}^{(N)} = \{00, 01, 01, 11, 11\}$:

$\mathbf{x} = (x_1, x_2)$	00	01	10	11
$\mathbb{P}(\mathbf{x})$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
$n_{\mathbf{x}}$	1	2	0	2
$\hat{\mathbb{P}}^{(N)}(\mathbf{x})$	$\frac{1}{5}$	$\frac{2}{5}$	0	$\frac{2}{5}$

Q 3.2.1 À l'aide de $Y_{\mathbf{x}} := (\mathbf{X} = \mathbf{x}) = (X_1 = x_1$ et $\ldots$ et $X_d = x_d)$, variable de Bernoulli qui vaut 1 si $\mathbf{X}$ prend la valeur $\mathbf{x}$, montrer que

$$\forall \mathbf{x} \in D \quad \lim_{N \to \infty} \hat{\mathbb{P}}^{(N)}(\mathbf{x}) = \mathbb{P}(\mathbf{x})$$

On utilisera les résultats de la question 1 avec la fonction $f_{\mathbf{x}}(\mathbf{x}')$ qui vaut 1 si $\mathbf{x} = \mathbf{x}'$ et 0 sinon.

Q 3.2.2 Montrer alors une seconde formulation du schéma de Monte-Carlo :

$$\mathbb{E}_{\mathbb{P}}(f) = \lim_{N \to \infty} \mathbb{E}_{\hat{\mathbb{P}}^{(N)}}(f) \tag{9}$$

On pourra montrer d'abord que $\mu_1^{(N)} = \mathbb{E}_{\hat{\mathbb{P}}^{(N)}}(f)$ (attention à bien différentier les sommes sur la base $\mathcal{X}^{(N)}$ et les sommes sur l'ensemble du domaine D).

Q 3.2.3 En termes de complexité, quelle approximation de $\mathbb{E}_{\mathbb{P}}(f)$ vous semble préférable entre les équations (8) et (9) ?

Q 3.3 Approximation de l'entropie $H(\mathbb{P})$ (bonus)
L'entropie d'une loi $\mathbb{P}$ est définie par :

$$H(\mathbb{P}) = -\mathbb{E}_{\mathbb{P}} \log \mathbb{P} = -\sum_{\mathbf{x} \in D} \mathbb{P}(\mathbf{x}) \cdot \log \mathbb{P}(\mathbf{x}) \tag{10}$$

Pour les mêmes raisons que l'équation (7), on ne peut pas utiliser cette formule pour calculer l'entropie quand d devient trop grand.

Q 3.3.1 Rappeler la définition de $LL(\mathcal{X}^{(N)} ; \mathbb{P})$: la log-vraisemblance de la base $\mathcal{X}^{(N)}$ pour la loi $\mathbb{P}$.
En déduire que le schéma de Monte-Carlo pour approcher l'entropie $H(\mathbb{P})$ revient à calculer $\overline{LL}(\mathcal{X}^{(N)} ; \mathbb{P})$: la log-vraisemblance moyenne sur $\mathcal{X}^{(N)}$ de la loi :

$$\overline{LL}(\mathcal{X}^{(N)} ; \mathbb{P}) = \frac{1}{N} LL(\mathcal{X}^{(N)} ; \mathbb{P})$$

Q 3.3.2 La *cross-entropy* (entropie croisée) de deux distributions $\mathbb{P}$ et $\mathbb{Q}$ est définie par

$$H(\mathbb{P}, \mathbb{Q}) = -\mathbb{E}_{\mathbb{P}} \log \mathbb{Q} = -\sum_{\mathbf{x} \in D} \mathbb{P}(\mathbf{x}) \cdot \log \mathbb{Q}(\mathbf{x}) \tag{11}$$

En utilisant la seconde caractérisation du schéma de Monte-Carlo, montrer qu'on peut approcher l'entropie de $\mathbb{P}$ par la *cross-entropy* de $\hat{\mathbb{P}}^{(N)}$ et $\mathbb{P}$.

$$H(\mathbb{P}) = \lim_{N \to \infty} H(\hat{\mathbb{P}}^{(N)}, \mathbb{P}) \tag{12}$$

Vérifier également que $H(\hat{\mathbb{P}}^{(N)}, \mathbb{P}) = -\overline{LL}(\mathcal{X}^{(N)} ; \mathbb{P})$

Q 3.3.3 Cette dernière équation permet de calculer une estimation de l'entropie de $\mathbb{P}$ à partir d'une base de donnée, pour peu que $\mathbb{P}$ soit assez bien connue pour pouvoir calculer facilement $\mathbb{P}(\mathbf{x})$ pour tout vecteur $\mathbf{x}$ de D.
Toutefois, souvent, $\mathbb{P}$ n'est pas assez bien connue et ne peut qu'être approchée par $\hat{\mathbb{P}}^{(N)}$. Dans ce cas, ni l'entropie de $\mathbb{P}$, ni la *cross-entropy* $H(\hat{\mathbb{P}}^{(N)}, \mathbb{P})$ ne sont calculables. L'entropie empirique $H(\hat{\mathbb{P}}^{(N)})$ est, elle, toujours calculable (pour des valeurs raisonnables de N).
Montrer en effet que le calcul de celle-ci ne dépend pas du modèle ($\mathbb{P}$) mais uniquement des comptages $n_{\mathbf{x}}$ dans la base $\mathcal{X}^{(N)}$.

Q 3.3.4 Par contre, l'estimateur obtenu ne vérifie pas trivialement les propriétés nécessaires pour être sans biais. Dans cette question, il s'agit de montrer que la différence entre la *cross-entropy* $H(\hat{\mathbb{P}}^{(N)}, \mathbb{P})$ et l'entropie empirique $H(\hat{\mathbb{P}}^{(N)})$ est asymptotiquement négligeable.
La divergence de Kullback-Leibler $D_{KL}(\mathbb{P}, \mathbb{Q})$ est une mesure de dissimilarité entre les distributions $\mathbb{P}$ et $\mathbb{Q}$:

$$D_{KL}(\mathbb{P}, \mathbb{Q}) = \sum_{\mathbf{x} \in D} \mathbb{P}(\mathbf{x}) \cdot \log \frac{\mathbb{P}(\mathbf{x})}{\mathbb{Q}(\mathbf{x})} = \mathbb{E}_{\mathbb{P}}(log \frac{\mathbb{P}}{\mathbb{Q}})$$

Elle possède de nombreuses propriétés intéressantes. En particulier,

$$D_{KL}(\mathbb{P}, \mathbb{Q}) \geq 0 \tag{13}$$

$$D_{KL}(\mathbb{P}, \mathbb{Q}) = 0 \iff \mathbb{P} = \mathbb{Q} \tag{14}$$

Montrer que $H(\hat{\mathbb{P}}^{(N)}, \mathbb{P}) - H(\hat{\mathbb{P}}^{(N)}) = D_{KL}(\hat{\mathbb{P}}^{(N)}, \mathbb{P})$
En conclure que $\lim_{N \to \infty} H(\hat{\mathbb{P}}^{(N)}, \mathbb{P}) - H(\hat{\mathbb{P}}^{(N)}) = 0$.
On peut donc utiliser l'entropie empirique comme estimateur sans biais de l'entropie de la loi $\mathbb{P}$.