APPRENTISSAGE STATISTIQUE AVANCÉ

EXAMEN - CONTRÔLE DE CONNAISSANCES (DURÉE 3 HEURES)

Les notes de cours ne sont pas autorisées, l'usage d'ordinateurs ou tablettes est prohibé.

OPTIMISATION (ELIMINATION DE VARIABLES)

La plupart des questions peuvent être traitées sans avoir résolu les précédentes.

Soit A une matrice de taille $n \times p$, soit b un vecteur-colonne de taille n et $\lambda > 0$. On s'intéresse au problème

$$\min_{x \in \mathbb{R}^p} \lambda \|x\|_1 + \frac{1}{2}\|Ax - b\|^2 \tag{1}$$

où $\|x\|_1 = \sum_{i=1}^p |x_i|$ si $x_1, \ldots, x_p$ représentent les composantes du vecteur x et où $\|\,.\,\|$ est la norme euclidienne usuelle.

1. En apprentissage statistique, quel nom est souvent donné au problème (1) ?
2. En introduisant artificiellement une variable auxilaire $z = Ax$, on peut réécrire le problème (1) sous la forme

$$\min_{\substack{x \in \mathbb{R}^p, z \in \mathbb{R}^n \\ Ax = z}} \lambda \|x\|_1 + \frac{1}{2}\|z - b\|^2 . \tag{2}$$

   Le couple (x, z) constitue la variable "primale". Ecrire le lagrangien $L((x, z); \nu)$ associé au problème (2), où ν est le multiplicateur de Lagrange associé à la contrainte $Ax - z = 0$.
3. On pose $f(x) = \|x\|_1$ et on note A^T la transposée de A. Justifier le fait qu'un couple $((x, z); \nu)$ est un point selle du lagrangien si et seulement si

$$\begin{aligned} 0 &\in \partial f(x) + \frac{A^T \nu}{\lambda} \\ \nu &= Ax - b \\ z &= Ax . \end{aligned} \tag{3}$$

   On admettra (ou on se souviendra) que $\partial f(x) = \mathrm{sign}(x_1) \times \cdots \times \mathrm{sign}(x_n)$. Autrement dit, un élément u satisfait $u \in \partial f(x)$ si et seulement si pour tout $i = 1, \ldots, p$, $u_i \in \mathrm{sign}(x_i)$. Cela revient encore à dire que pour tout i,

$$\begin{aligned} u_i &= 1 \quad && \text{si } x_i > 0 \\ u_i &= -1 \quad && \text{si } x_i < 0 \\ u_i &\in [-1, 1] \quad && \text{si } x_i = 0 . \end{aligned}$$

   On note $A = (a_1, \ldots, a_p)$ où $a_1, \ldots, a_p$ sont les colonnes de la matrice A. On rappelle que la $i$ème composante du vecteur $A^T \nu$ vaut $a_i^T \nu$.
4. Soit $((x, z); \nu)$ un point selle du lagrangien. En utilisant l'équation (3), montrer que pour tout $i = 1, \ldots, p$,

$$|a_i^T \nu| < \lambda \implies x_i = 0 .$$

5. On pose $\lambda_0 = \max_{i=1\dots p} |a_i^T b|$. En utilisant la question 3, vérifier que $((0,0);-b)$ est un point selle du lagrangien. En déduire que le point $x = 0$ est solution du problème (1).
6. On suppose dorénavant que $\lambda < \lambda_0$. On note Φ la fonction duale, c'est à dire $\Phi(\nu) = \inf_{x,z} L((x,z);\nu)$. On rappelle qu'une solution duale est un maximiseur de Φ. On pose

$$G(\nu) = -\frac{\|\nu + b\|^2}{2} + \frac{\|b\|^2}{2}.$$

On **admettra** le résultat suivant :

$$\Phi(\nu) = \begin{cases} G(\nu) & \text{si } \forall i, |a_i^T \nu| \leq \lambda \\ -\infty & \text{sinon.} \end{cases}$$

On pose $\gamma = \Phi(-\frac{\lambda}{\lambda_0} b)$. Montrer que $\gamma > -\infty$. Exprimer γ en fonction de λ, λ_0 et b.

7. Justifier que toute solution duale ν^* appartient à l'ensemble

$$S = \{\nu \in \mathbb{R}^n \,:\, G(\nu) \geq \gamma \text{ et } \forall i,\, |a_i^T \nu| \leq \lambda\}.$$

8. On fixe dorénavant un indice $i \in \{1,\dots,p\}$. On pose $T_i = \max_{\nu \in S} |a_i^T \nu|$. En utilisant la question 4, montrer que si $T_i < \lambda$ alors toute solution x du problème (1) satisfait $x_i = 0$.
9. Un calcul direct (que l'on ne demande pas d'effectuer) montre que

$$T_i = |a_i^T b| + \|a_i\| \sqrt{\|b\|^2 - 2\gamma}. \tag{4}$$

Avant de résoudre numériquement le problème (1), un ingénieur décide, pour chaque indice i, de comparer la quantité (4) au seuil λ. Quel intérêt pratique ce test peut-il avoir ?

RÉSEAUX DE NEURONES

1. En quoi l'optimisation d'un réseau de neurones est-elle plus difficile que celle d'un modèle linéaire ?
2. Quel est le coût d'une itération de gradient stochastique ? De gradient "batch" ? Lequel converge plus vite au début ? A la fin ?

MODÈLES DE MARKOV CACHÉS

Génération d'une suite d'états d'un modèle de Markov. On considère un modèle de Markov *ergodique* (toutes les transitions entre états sont autorisées). Le nombre états dans ce modèle est $Q = 2$. On donne le vecteur des probabilités initiales d'états :

$$\pi = \begin{bmatrix} 0.35 & 0.65 \end{bmatrix}$$

et la matrice de transitions entre états :

$$A = \begin{bmatrix} 0.35 & 0.65 \\ 0.2 & 0.8 \end{bmatrix}$$

On veut générer une séquence d'états $q_1, q_2, \dots q_T$ de longueur $T = 5$ correspondant à ce modèle.

On commence par générer le premier état suivant la loi donnée par le vecteur π, puis on génère les quatre états suivants en utilisant la matrice A.

1. Comme nous l'avons vu en TP, pour générer des états, on tire des nombres aléatoires suivant une loi uniforme entre 0 et 1. Expliquer le principe du tirage pour générer des états.

2. Les 5 nombres tirés sont les suivants :
   $u_1 = 0.92$ (pour l'état q_1) et
   $u_2 = 0.31, u_3 = 0.1, u_4 = 0.4, u_5 = 0.01$ (pour les états q_2 à q_5)
   Quelle est la suite d'états correspondant aux tirages des $u_i, i = 1, .., 5$ (justifier votre réponse) ?

CHAMPS DE MARKOV CONDITIONNELS ET SORTIES STRUCTURÉES

1. Quelle est la différence entre caractéristiques (features) et fonctions de caractéristiques (feature functions) ? Quel est l'intérêt de ces fonctions de caractéristiques ?
2. Quels sont les points forts des CRF par rapport aux HMM pour la classification de données séquentielles ?

MÉTHODES À NOYAUX AVANCÉES

1. Définir le problème de l'apprentissage statistique semi-supervisé.
2. Proposer une fonction de coût appropriée pour la régression semi-supervisée dans un espace de Hilbert à noyau autoreproduisant $\mathcal{H}_k$ défini à partir d'un noyau positif défini k sur $\mathbb{R}^p$. Expliquer chaque terme de cette fonction de coût.
3. Sous quelle forme s'écrit la solution de ce problème ?
4. Donner deux méthodes de résolution de ce problème d'optimisation.

APPRENTISSAGE PAR RENFORCEMENT

1. Présenter en le justifiant le bandit "ϵ-greedy"
2. Rappeler le critère Upper Confidence Bound (UCB) et expliquer pourquoi UCB permet l'exploration.