Nom de l'étudiant: _____

FACULTE DES ARTS ET DES SCIENCES
DEPARTEMENT D'INFORMATIQUE ET DE RECHERCHE OPERATIONNELLE

TITRE DU COURS: **Algorithmes d'apprentissage**
SIGLE DU COURS : **IFT6266 A04**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A04: 26 octobre 2003 HEURE: 13h - 15h
SALLE: PAA-3195

DIRECTIVES PEDAGOGIQUES: - Une feuille recto-verso permise
- Répondre directement sur le questionnaire.
- Soyez brefs et précis dans vos réponses.
- Si vous manquez de temps: l'important est de montrer que vous avez compris le problème, plutôt que les détails de la réponse.

---

1. Vous allez appliquer l'algorithme EM à une version simple du mélange d'experts. Les variables observées sont $X \in \mathbf{R}^n$ et $Y \in \mathbf{R}$. La variable cachée est $Z \in \{1 \dots m\}$, discrète. La variable X sert seulement à conditionner la distribution de Y (on ne s'intérese pas à la densité de X). Le modèle permet de représenter $P_Y(y|X)$ à partir de la jointe $P(Y = y, Z = i|X) = P_Y(y|Z = i, X)P(Z = i|X)$. Le modèle pour chaque expert i est $P_Y(y|Z = i, X = x)$ une densité normale d'espérance $\mu_i(x) = w_i \cdot x$ et de variance σ_i^2 indépendante de x. Les paramètres sont $\theta = (w_1, ...w_m, \sigma_1, ..., \sigma_m)$ avec $w_i \in \mathbf{R}^n$. Pour simplifier on va supposer que $P(Z = i|X = x)$ est une fonction donnée et fixe de x. S'il vous plaît utilisez la notation $P(Z = i|X = x_t) = p_{it}$ pour simplifier la correction. Soit $D = \{(x_t, y_t)\}$ l'ensemble d'apprentissage.

   (a) Montrez comment calculer le postérieur $\pi_{ti} = P(Z = i|Y = y_t, X = x_t)$.

   (b) Développer la formule de la fonction auxiliaire (étape E de l'algorithme) $Q(\theta, \theta') = \sum_t E_Z[\log P_Y(y_t, Z|X = x_t, \theta)|Y = y_t, X = x_t, \theta']$ et calculer sa dérivée par rapport aux w_i et σ_i^{-1}.

   (c) Utilisez ces dérivées pour obtenir la formule de re-estimation des paramètres dans l'étape M de l'algorithme.

2. Nous avons vu comment utiliser une méthode non-paramétrique pour estimer l'espérance conditionnelle d'une variable aléatoire $Y \in \mathbf{R}^d$ étant donné une variable aléatoire X (e.g. la "fenêtre de Parzen": $f(x) = \frac{\sum_t y_t K_D(x,x_t)}{\sum_t K_D(x,x_t)}$ où $D = \{(x_t, y_t)\}$). Appliquer le même principe que dans la fenêtre de Parzen mais pour estimer la **matrice de covariance conditionnelle** de Y ($d > 1$) étant donné X. Donner un algorithme ou une formule pour cet estimateur.

   QUESTION BONI: calculez le biais de cet estimateur pour le K_D du K-plus-proche-voisin, c'est à dire, $K_D(x, y) = 1$ si y est un des k plus proches voisins de x dans D, 0 autrement.

3. Soit un algorithme d'apprentissage qui obéit au principe de minimisation de l'erreur d'apprentissage, i.e. on choisit une fonction $f \in \mathcal{F}$ qui minimise l'erreur moyenne $e(f, D) = \frac{1}{n} \sum_{i=1}^{n} C(f, z_i)$ sur les exemples d'apprentissage $D = \{z_i\}$. Soit $f^*(D) = \text{argmin}_{\text{f}\in\mathcal{F}}\text{e}(\text{f}, \text{D})$. On suppose z_i tirés i.i.d. d'une loi inconnue P. Soit D_2 un deuxième ensemble de données tirées indépendamment de D selon la même loi.

   (a) Est-ce qu'on s'attend à trouver que $e(f^*(D), D_2)$ est inférieur, égal, ou supérieur à $e(f^*(D), D)$ (en espérance sur D et D_2)?
   Pourquoi selon vous? (je ne vous demande pas un argument mathématique complet, seulement une intuition ou un exemple qui explique votre réponse).

(b) Si on augmente la richesse (complexité, capacité, ...) de la classe de fonction $\mathcal{F}$, qu'est-ce qu'on peut dire sur l'effet que cela aurait sur $e(f^*(D), D)$ et $e(f^*(D), D_2)$? Pourquoi selon vous?

(c) Nommez 3 *hyper-paramètres* pour un algorithme d'apprentissage de réseau de neurones multi-couche.

4. Considérez un réseau de neurones qui prédit $f(x) = v \cdot \tanh(Wx)$ et minimise la moyenne des erreurs quadratiques $(f(x) - y)^2$, avec paramètres $v \in \mathbf{R}^m$ et $W \in \mathbf{R}^{m \times n}$, et $x \in \mathbf{R}^n$. Montrez que si on initialise le modèle avec $W = 0$ et $v = 0$ les poids ne bougent pas si on fait de la descente de gradient.

5. Une caractéristique des modèles paramétriques de statistique classique est que le nombre de paramètres (et la complexité de la fonction résultante) ne dépend pas du nombre d'exemples. Par contre avec les modèles non-paramétriques on choisit un hyper-paramètre (par exemple la taille du voisinage dans les K-plus-proches-voisins) pour adapter la complexité de la fonction au nombre d'exemples. Dans ce contexte, les réseaux de neurones et l'estimation de densité par mélange de Gaussiennes peuvent être vus comme paramétriques ou comme non-paramétriques. Expliquer ce qui fait qu'on devrait soit les considérer comme paramétriques ou comme non-paramétriques.