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 A03**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN FINAL A03: 9 décembre 2003 HEURE: 15h30 - 17h25
SALLE: Z-337

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. Supposons un ensemble d'apprentissage supervisé $D = \{(x_i, y_i)\}_{i=1}^n$ avec les entrées $x_i \in \mathbf{R}^d$, et un algorithme d'apprentissage A qui prend un ensemble de données et retourne une fonction. Si on **rajoute une variable d'entrée supplémentaire** (i.e. on incrémente d, en gardant n fixe), que pouvez-vous dire sur ce qui peut arriver à

   (a) l'erreur d'apprentissage ?

   (b) l'erreur de généralisation ?

   (c) la variance de $f = A(D)$?

   (d) le biais de $f = A(D)$?

2. Si on vous donne un ensemble de données D avec $n = 100$ exemples et la dimension de chaque vecteur d'entrée est $d = 1000$, et que vous observez qu'une régression linéaire ordinaire généralise très mal, suggérez plusieurs méthodes à essayer pour améliorer la généralisation (faites référence à des techniques vues en classe, pas besoin de donner le détail algorithmique ou mathématique, juste nommez-les).

3. Si le nombre d'exemples d'apprentissage est $n = 100$, pensez-vous que l'apprentissage d'un réseau de neurones serait plus rapide par descente de gradient stochastique ou par gradients conjugué? justifiez brièvement votre réponse.

4. Expliquez dans quelle circonstance il vaut mieux utiliser l'une ou l'autre des procédures d'estimation de l'erreur de généralisation suivantes:

   (a) séparation des données en 2 parties: entraînement + test.

   (b) validation croisée par bloc (*k-fold cross-validation*).

   (c) validation séquentielle.

5. Montrez que l'algorithme de clustering des k-moyennes (à chaque itération on déplace le prototype μ_k de chaque groupe au centre de masse des exemples x_i qui en sont le plus proche) converge vers une solution qui est un **minimum** (possiblement local) de l'erreur de reconstruction

$$C = \sum_i ||x_i - \mu_{p(x_i)}||^2$$

où $p(x_i)$ est le numéro du prototype le plus proche de x_i:

$$p(x_i) = \text{argmin}_k ||x_i - \mu_k||^2.$$

N.B. La dérivée de $p(x_i)$ par rapport à μ_k peut être considérée nulle (elle l'est presque partout). Quelles sont les conditions pour obtenir un minimum?

6. On peut généraliser le principe des mélanges de distributions à des mélanges de distributions conditionnelles. Par exemple, on peut représenter $P(Y=y|X=x)$ par
$$P(Y=y|X=x)=\sum_h P(Y=y|H=h,X=x)P(H=h|X=x).$$
L'ensemble de données contient des paires (x_i, y_i). La variable H est donc cachée. Supposons que les modèles $P(H|X)$ et $P(Y|H,X)$ soient faciles à optimiser si H était connue dans les données d'apprentissage. Expliquer comment utiliser l'algorithme EM pour entraîner un tel modèle. Vous pouvez supposer qu'on vous fournit une procédure pour choisir les paramètres de ces distributions **SI** H **ÉTAIT CONNU** (i.e. si les données étaient des tuples (x_i, h_i, y_i) cette procédure permet de maximiser la log-vraisemblance **pondérée**
$$\sum_i w_i \log P_\alpha(Y=y_i|H=h_i,X=x_i)$$
en α, ou bien elle peut maximiser
$$\sum_i w_i \log P_\beta(H=h_i|X=x_i).$$
en β. Les w_i sont des poids arbitraires que vous pouvez choisir grâce à EM. N.B. On appelle ce type de modèle un *mélange d'experts* (mixture of experts). Commencez par énoncer l'algorithme EM en général et ensuite trouvez les expressions particulières dont vous avez besoin ici.

7. Considérez la version Bayesienne de la régression linéaire avec régularisation “ridge” avec coefficient λ. Comment nomme-t-on la loi qui donne la distribution à priori sur les paramètres? A quoi correspond le coefficient λ dans cette distribution?.

8. Dans le même contexte, exprimer (sans essayer de calculer les intégrales explicitement) la prédiction qui serait effectuée avec la version Bayesienne de la régression linéaire "ridge", **étant donné** un vecteur d'entrée x, un ensemble de données $D = \{(x_i, y_i)\}$, l'hypothèse de bruit Gaussien, et l'hypothèse ci-haut sur la distribution à priori des paramètres?

9. Donnez le pseudo-code pour un algorithme de réduction de dimensionalité **non-linéaire**, c'est à dire un algorithme d'apprentissage non-supervisé qui permet de transformer les données (e.g. un ensemble de x_i) de vecteurs dans $\mathbf{R}^d$ en vecteurs z_i dans $\mathbf{R}^{d'}$ avec $d' < d$ tout en préservant le plus d'information (nous en avons vus plusieurs en classe ou bien vous pouvez en inventer un à partir des connaissances apprises en classe). **Non-linéaire** veut dire que chaque z_i n'est pas obtenu simplement par une transformation affine de x_i, comme dans l'Analyse en Composante Principale.