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 A05**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN FINAL A05: 8 décembre 2005 HEURE: 11h30 - 13h30
SALLE: PAA-3195

DIRECTIVES PÉDAGOGIQUES:
- Livre du cours et notes de cours permis (tout ce qui est sur le site du cours).
- Répondre directement sur le questionnaire. Vous pouvez utiliser l'arrière des pages aussi si vous en avez besoin.
- 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.**
- Échanger des informations lors d'un examen (ou autres formes de tricherie) est du **plagiat**, qui est passible de sanctions allant jusqu'à l'exclusion du programme.

---

1. Soient X, Y deux variables aléatoires de loi jointe $P(X, Y)$ inconnue, avec $X \in \mathcal{X}$, $Y \in \mathcal{Y}$. Soit $D_n = \{(x_i, y_i)\}_{i=1}^n$ un ensemble de n exemples tirés de P. Soit A un algorithme d'apprentissage, i.e. une fonctionnelle de D_n vers une fonction $f : \mathcal{X} \to \mathcal{Y}$. Soit $Q : \mathcal{Y} \times \mathcal{Y}$ une fonction de coût, avec $Q(f(X), Y)$ la perte encourue quand on prédit $f(X)$ alors que Y est réalisée.

   En utilisant la notation ci-haut, définir

   (a) Erreur moyenne d'apprentissage.

   (b) Erreur de généralisation.

(c) Critère d'apprentissage C **avec terme de régularisation** et coefficient de régularisation λ.

(d) Si on peut minimiser C analytiquement, quelle **équation** doit-on résoudre et quelle **inégalité** doit-on vérifier?

(e) Au fur et à mesure qu'on augmente λ, quel sera l'effet sur l'erreur d'apprentissage? et sur l'erreur de généralisation? (considérez l'intervalle $\lambda \in [0, \infty)$).

(f) Si n augmente, quel serait l'effet probable sur l'erreur d'apprentissage? sur l'erreur de généralisation?

(g) Pour n fixe, il existe une valeur optimale de λ (en terme d'erreur de généralisation), que l'on écrira $\lambda^*(n)$. Quand n augmente, est-ce que $\lambda^*(n)$ augmente ou diminue? **expliquez votre réponse**.

2. Soit A un algorithme d'apprentissage qui prend en argument un ensemble de données non-étiquetées et produit une fonction de densité. Par exemple $U = \{x_1, \ldots, x_m\}$ et $f = A(U)$ est une densité, i.e. $\int f(x)dx = 1$ et $f(x) \geq 0$. Supposons que l'on vous donne un ensemble de données D_n défini comme dans la question 1, avec $\mathcal{Y} \in \{1, 2, \ldots, C\}$ un petit ensemble de C classes. Montrez comment on peut utiliser l'algorithme A pour faire de la classification probabiliste, i.e., estimer $P(Y|X)$: pour cela écrivez le pseudo-code d'un algorithme qui prend D_n en entrée et produit une fonction q telle que $q(x, y)$ estime $P(Y = y|X = x)$.

3. Vous allez comparer deux critères pour l'entraînement d'un réseau de neurone en classification probabiliste. On a des données $D_n = \{(x_t, y_t)\}_{t=1}^n$ avec $y_t \in \{0, 1\}$. Dans les deux cas on estime $P(Y = 1|X = x)$ par une fonction $f(x) = \text{sigmoid}(a(x)) = \hat{P}(Y = 1|X = x)$, avec $a(x) = b + W \tanh(c + Vx)$, de paramètre $\theta = (b, W, c, V)$. Mais dans le premier cas on utilise le critère de log-vraisemblance $C = \sum_t C_t = -\sum_t \log \hat{P}(Y = y_t|X = x_t)$ alors que dans le deuxième cas on utilise le critère d'erreur quadratique $C = \sum_t C_t = \sum_t (f(x) - y_t)^2$.

   (a) Montrez que dans les deux cas, quand $n \to \infty$ on obtient que $f(x) \to P(Y = 1|x)$. Pour simplifier les mathématiques, supposez que l'ensemble des valeurs que X peut prendre est énumérable.

(b) Montrez que quand le neurone de sortie sature ($f(x_t)$ approche 1 ou 0) les gradients $\frac{\partial C_t}{\partial \theta}$ s'évanouissent (approchent 0) dans le cas du critère quadratique mais ne s'évanouissent pas nécessairement dans le cas du critère de log-vraisemblance.

4. Dans l'algorithme MDS (*Multi-Dimensional Scaling*) on convertit des supposées distances d_{ij} données (qu'on suppose correspondre à des positions x_i inconnuees telles que $d_{ij} = ||x_i - x_j||$) en produits scalaires B_{ij} (c'est à dire qu'on devrait avoir $B_{ij} = x_i \cdot x_j$). Soient $\mu_i = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2$ et $\mu = \frac{1}{n}\sum_{i=1}^{n} \mu_i$ les moyennes de distances.

   Prouvez que cette formule de transformation, $B_{ij} = -0.5(d_{ij}^2 - \mu_i - \mu_j + \mu)$ est correcte, i.e. que l'on obtient bien les produits scalaires (à une transformation rigide près des coordonées x_i).

5. On vous donne un échantillon de données $\{x_1, \ldots, x_n\}$ avec les x_i des observations de la variable aléatoire $X \in \mathbf{R}^D$. Supposons que l'on vous donne aussi un algorithme A qui identifie une variété affine de dimension d dans $\mathbf{R}^D$, variété où la densité se concentre. Cette variété est spécifiée par un point $\mu \in \mathbf{R}^D$ sur la variété et par un ensemble de d vecteurs de base $v_i \in \mathbf{R}^D$ tels que tout point $\hat{x}$ sur la variété peut s'écrire sous la forme $\hat{x} = \mu + \sum_{i=1}^{d} \alpha_i v_i$. Comme dans l'analyse en composante principale, on peut choisir pour un x donné qui n'est pas sur la variété un $\hat{x}$ sur la variété et qui soit le plus proche de x. Comment, à partir d'un estimateur de la densité sur la variété (donc des $\alpha \in \mathbf{R}^d$, coordonnées réduites des exemples), pourrait-on construire un estimateur de la densité pour $X \in \mathbf{R}^D$? On va supposer que les écarts entre un x et sa projection $\hat{x}$ sur la variété peuvent être modélisés par une simple normale de variance σ^2 (dans toutes les directions orthogonales à la variété). Donnez **les grandes lignes** d'un algorithme qui estime la densité des x de la manière suivante: d'abord utiliser A pour trouver la variété, ensuite estimer σ^2 (pour la densité des variations hors-variété) et estimer la densité sur la variété. Comment combiner les trois choses pour obtenir un estimateur de densité pour les x dans $\mathbf{R}^D$?

   (a) Dans un premier temps supposez que la densité sur la variété (donc celle des coordonées réduites des données sur la variété) peut être modélisée par une Gaussienne.

   (b) Proposez un moyen de généraliser la solution (a) de façon à pouvoir modéliser une distribution non-Gaussienne sur la variété (en gardant le bruit orthogonal à la variété Gaussien).