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 A06**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A06: 19 octobre 2006 HEURE: 13h - 14h50
SALLE: PAA-1411

DIRECTIVES PÉDAGOGIQUES:
- Documentation permise.
- 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.
- Suggestion: commencez par les questions les plus faciles (non **BONUS**)

1. Considérons le cas "standard" d'un algorithme d'apprentissage qui consiste en la minimisation d'un critère d'apprentissage régularisé C (erreur d'apprentissage plus terme de régularisation pénalisant la complexité), en choisissant le prédicteur f parmi une classe $\mathcal{F}$ de fonctions paramétrisées par un vecteur de paramètres θ. Soit n le nombre d'exemples d'apprentissage, et λ le coefficient de régularisation (qui contrôle la force de la pénalité sur la complexité de la solution). On minimise C par une méthode quelconque d'optimisation Soit $A(D)$ la fonction estimée avec des données D par l'algorithme d'apprentissage A. Les données sont des paires (x, y).

   POUR CETTE QUESTION ON ATTEND DES RÉPONSES COURTES!

   (a) Comment est-ce que l'on s'attend à ce que la différence entre l'erreur de généralisation et l'erreur d'apprentissage moyenne varie, au fur et à mesure que n augmente? pourquoi?

   (b) Comment est-ce que la valeur de λ affecte la richesse de la classe des fonctions que l'on peut obtenir en solution? (i.e. effet d'augmenter ou diminuer λ)

(c) Comment est-ce que le nombre d'itération de l'optimisation affecte la richesse de la classe des fonctions que l'on peut obtenir en solution?

(d) Dans le cas de la régression, la réponse optimale est $f_{opt}(x) = E[Y|X]$. De quelle manière est-ce que l'on s'attend que le *biais* de l'estimateur (le carré de la moyenne sur D des différences entre $A(D)(x)$ et $f_{opt}(x)$) varie en fonction de la richesse des fonctions obtenues? en fonction de n?

(e) Dans le même contexte, de quelle manière est-ce que l'on s'attend que la *variance* de l'estimateur (les variations de $A(D)(x)$ quand on varie D) change en fonction de la richesse des fonctions obtenues? en fonction de n?

(f) (**BONUS**) Pour toute loi *a priori* sur θ, la maximisation Bayesienne de $P(\theta|D)$ est équivalente à notre critère de régularisation C avec une certaine forme de régularisation. La densité de Laplace est $p(z) \propto e^{-|z|/\sigma}$ pour une v.a. scalaire z. Si on l'utilise comme loi *a priori* sur les paramètres θ, en supposant les θ_i indépendents entre eux sous cette loi, à quel terme de régularisation est-ce que cela donne lieu?

2. Considérez l'estimateur de l'erreur de généralisation donné par la validation croisée $kfold$ (à k partitions) sur un ensemble de données D avec n exemples. Montrez qu'il estime de façon non-biaisée (donc avec erreur 0 en espérance sur le choix de D) l'erreur de généralisation quand on entraîne avec $\frac{n(k-1)}{k}$ exemples.

3. Vous allez considérez des problèmes où l'on obtient 0 erreur d'apprentissage.

   (a) Montrez qu'avec $n-1$ variables d'entrée (et n exemples), la régression linéaire donne une solution "parfaite" (0 erreur) en terme d'erreur d'apprentissage. Il y a des cas pathologiques et improbables où cela ne fonctionne pas: lesquels? (**BONUS**): Donnez une réponse qui inclut aussi le cas général où le nombre d'entrées est d avec $d \geq n$.

(b) Considérez un réseau de neurones avec m unités cachées, et $m \geq n$ (n le nombre d'exemples), pour la régression. Montrez que l'on peut presque toujours apprendre par coeur dans ce cas. Il y a une condition bénigne sur les poids des unités cachées, qui sera vraie presque tout le temps même avec une initialisation aléatoire. Vous pouvez supposer vrai le théorème impliqué par la question précédente. (**BONUS**): Qu'arrive-t-il si le nombre d'entrées d est plus petit que m? Dans quel cas cela pourrait-il poser un problème?

4. (**BONUS**) Considérez l'algorithme itératif suivant pour résoudre un système d'équations linéaires, i.e., résoudre $Ax = b$, avec A carrée et inversible. On part d'une solution x_0 et on itère $x_t = x_{t-1} + \epsilon(b - Ax_{t-1})$, avec ϵ un petit scalaire. Si on fait graduellement décroître ϵ, montrez pourquoi cette méthode converge vers la solution. Vous pouvez exploiter le fait que la descente de gradient converge.