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 A00**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A00: 14 novembre 2000 HEURE: 10h30 - 12h25
SALLE: PAA-1411

DIRECTIVES PEDAGOGIQUES: - Aucune documentation permise
- Répondre directement sur le questionnaire.
- Soyez brefs et précis dans vos réponses.

1. La minimisation du risque empirique peut parfois donner comme solution une fonction f qui généralise très mal.

    (a) Pourquoi?

    (b) Donnez un exemple où cela pourrait arriver même si f est choisie dans un ensemble paramétrisé par un petit nombre de paramètres (beaucoup plus petit que le nombre d'exemples) et même si f donne une erreur d'apprentissage très faible.

2. Dans le contexte de l'apprentissage par minimisation du risque empirique,

   (a) quelle est la relation entre la *capacité* h d'une classe de fonctions $\mathcal{F}$ et (1) l'erreur d'apprentissage, et (2) l'erreur de généralisation?

   (b) même question pour l'influence du nombre d'exemples d'apprentissage sur ces deux erreurs.

3. Quelle différence(s) y a-t-il entre des modèles d'apprentissage dits *non-paramétriques* et ceux dits *paramétriques*?

4. La méthode des histogrammes a le défaut de produire des solutions non-lisses. Quelle modification pourriez-vous y faire pour éviter cela?

5. Nous avons introduit une méthode de classification supervisée qui permet d'obtenir une très faible erreur d'apprentissage tout en généralisant assez bien (sans "overfitter"). Quel est le nom et le principe de base de cette méthode?

6. Nous avons montré au TP2 que la pénalisation de la norme des poids dans une régression linéaire peut toujours améliorer la généralisation. De manière similaire, prouvez qu'on peut toujours améliorer l'estimation d'une espérance en divisant la moyenne par une constante > 1. On suppose que les données z_t $(t = 1..T)$ sont générées i.i.d. avec une espérance μ et une variance $\sigma^2 > 0$. L'estimateur par "shrinkage" est donc

$$\hat{\mu}_\lambda = \arg\min_m \frac{1}{T}\sum_{t=1}^{T}(m - z_t)^2 + \lambda m^2 = \frac{\frac{1}{T}\sum_t z_t}{(1+\lambda)}$$

où λ joue le rôle du coefficient de pénalisation ("weight decay"). N.B. Similairement au TP2 on peut montrer que le λ "optimal" en espérance (i.e. pas pour un ensemble fixe de données) est $\frac{\sigma^2}{T\mu^2}$.

7. Sachant qu'un réseau de neurones peut représenter une distribution conditionnelle $P(Y|X)$ (par exemple si Y est une v.a. discrète on peut utiliser sur la dernière couche la fonction softmax), comment pourrait-on utiliser des réseaux de neurones pour représenter la distribution jointe de n variables aléatoires discrètes $(Z_1...Z_n)$? Quel avantage potentiel y voyez-vous par rapport à un modèle qui assigne à chaque combinaison possible des valeurs de ces variables un paramètre libre (qui représente la probabilité de cette combinaision)?

8. Un modèle de Markov caché (HMM) avec N valeurs de l'état caché permet de représenter la distribution jointe des observations y_t dans une séquence y_1^T de manière efficace:
   (a) La représentation est efficace en terme du nombre de paramètres. Pourquoi?
   (b) La représentation est efficace car elle permet le calcul de la vraisemblance efficacement (en temps linéaire plutôt qu'exponentiel par rapport à la longueur de la séquence). Expliquez pourquoi.

9. Comment pourrait-on diminuer la capacité d'une régression linéaire?

10. Comment pourrait-on augmenter la capacité d'un classifieur linéaire?

11. Considérez deux vecteurs i.i.d. $x, y \in \mathcal{R}^n$. Soit $d(x,y) = \frac{1}{n}\sum_{k=1}^{n}(x_k - y_k)^2$ la distance normalisée (par la dimension n) entre cette paire aléatoire de vecteurs. Montrez ce qui arrive à cette distance normalisée quand la dimension devient grande, i.e. $n \to \infty$ (suggestion: théorème limite centrale).