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 A98**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A98: 3 novembre 1998 HEURE: 8h30 - 10h29
SALLE: Z-310

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

1. Définir formellement le concept d'erreur de généralisation, d'erreur d'apprentissage et de minimisation du risque empirique.

2. Considérons un classifieur entraîné sur un ensemble fini d'exemples d'apprentissage. Sa réponse hors de l'échantillon d'apprentissage à un point (tiré de la même distribution d'où proviennent les données) est fondamentalement incertaine. Pour quelle(s) raison(s)?

3. Expliquez comment on peut utiliser le théorème d'inversion de Bayes pour construire un classifieur à partir d'un algorithme d'apprentissage qui fait de l'estimation de densité.

4. Donnez un ensemble d'exemples d'apprentissage pour la classification qui soit non-linéairement séparable.

5. La notion de capacité s'applique à des ensembles de fonctions. D'un autre côté, nous avons défini un principe différent du principe de minimisation du risque empirique, qu'on a appelé **régularisation**, où l'algorithme d'apprentissage cherche à minimiser la somme du risque empirique et d'un terme de régularisation qui dépend de la fonction choisie. Considérons le cas particulier où la classe de fonctions $\mathcal{F}$ a une forme paramétrique avec un vecteur de paramètres θ, et le terme de régularisation est de la forme $\lambda \sum_i \theta_i^2$, où λ est un "hyper-paramètre" dont on dit qu'il "contrôle la capacité". Pourquoi peut-on dire qu'il contrôle la capacité? Hint: pour répondre à cette question il faudra vous rapporter au fait que la capacité est une mesure qui s'applique à un *ensemble de fonctions.*

6. Expliquez le principe de l'algorithme E-M (Expectation-Maximization).

7. Essayez de démontrer que la décomposition biais-variance-bruit est correcte:

$$
\begin{aligned}
E_{D,X,Y}[(f^*(D)(X) - Y)^2] \quad = \quad & E_{X,Y}[(E[Y|X] - Y)^2] && \text{“bruit”} \\
& + E_X[(E_D[f^*(D)(X)] - E[Y|X])^2] && \text{“biais”} \\
& + E_{D,X}[(f^*(D)(X) - E_D[f^*(D)(X)])^2] && \text{“variance”}
\end{aligned}
$$

où D est un ensemble de données (x, y) tirées d'une distribution $P(X, Y)$, $f^*(D)$ est une fonction choisie selon D dans un ensemble $\mathcal{F}$ qui ne contient pas nécessairement l'espérance conditionnelle $f(x) = E[Y|x]$.

8. Pourquoi est-ce qu'il pourrait être incorrect d'appliquer l'algorithme de validation croisée par *K-fold cross-validation* pour estimer l'erreur de généralisation quand les données ne sont pas i.i.d.? Pour répondre à cette question, il vous suffit de montrer un exemple où on obtiendrait un estimé très biaisé de l'erreur de généralisation.

9. Le principe du maximum de vraisemblance peut parfois donner des résultats aberrants.

   (a) Donnez un exemple où cela peut arriver et expliquez la source du problème.

   (b) Proposez une solution, sous la forme d'un autre critère d'apprentissage (autre que la log-vraisemblance moyenne).

   (c) Montrez une application de cette solution au cas des mixtures de Gaussiennes.