Nom de l'étudiant: ______________________________

FACULTE DES ARTS ET DES SCIENCES
DEPARTEMENT D'INFORMATIQUE ET DE RECHERCHE OPERATIONNELLE

TITRE DU COURS: **Reseaux de Neurones et Apprentissage**
SIGLE DU COURS : **IFT6265 A96**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A96: 15 Octobre 1996 HEURE: 15h30 - 17h20
SALLE: Z-245

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

1. Expliquer brièvement (ou nommez) un algorithme d'apprentissage qui est garanti d'obtenir zéro erreur sur les exemples d'apprentissage. Quelle est sa capacité? Que peut-on dire sur l'erreur espérée sur de nouveaux exemples? Pourquoi?

2. Donnez un exemple de connaissance à priori concernant une tâche à apprendre, et comment cette connaissance est incorporée dans un algorithme d'apprentissage (ce qui permet de combiner deux sources d'information: les exemples et les connaissances à priori).

3. Construisez une porte logique NON (négation logique) avec un neurone formel (donnez son architecture et la valeur de ses paramètres).

4. Donnez un exemple de fonction qu'un neurone formel ne peut représenter.

5. Que veut-on dire quand on dit que l'algorithme **A** est "plus **discriminant**" que l'algorithme **B**?

6. L'erreur de généralisation peut être décomposée en un terme de "biais" et un terme de "variance".

   (a) Expliquez ces deux termes de manière qualitative (et si vous voulez, aussi mathématiquement, mais je voudrais surtout votre explication intuitive sur leur signification, par seulement des équations).

   (b) Qu'arrive-t-il à ces deux termes ainsi qu'à la capacité effective quand le nombre d'époques d'apprentissage (i.e. le temps de calcul disponible) augmente (pour un algorithme d'apprentissage itératif tel que la descente de gradient)?

7. Considèrez le problème d'apprentissage suivant (appelé modèle auto-regressif linéaire d'ordre deux). Il faut trouver les w_i qui minimisent l'erreur quadratique $C = 0.5 \sum_{t=2}^{L} (y_t - \hat{y}_t)^2$ avec le prédicteur linéaire $\hat{y}_t = w_0 + w_1 y_{t-1} + w_2 y_{t-2}$ où $w_i, y_t, \hat{y}_t \in R$.

   Dessinez le diagramme du flot des calculs (i.e. des dépendences entre les variables $C, y_t, \hat{y}_t, w_i$). Notez que les w_i sont partagés dans le temps. A partir du diagramme, montrez comment calculer le gradient $\frac{\partial C}{\partial w_i}$ par rapport aux paramètres. Peut-on minimiser C analytiquement?

8. Supposez que vous êtes dans la situation suivante: vous avez entrainé un réseau de neurones multi-couches avec 40 entrées, 30 unités cachées, et 3 sorties (pour un problème de régression ou de classification). Vous utilisez la descente de gradient stochastique pour l'apprentissage. Il y a N exemples disponibles.

   (a) Combien y-a-t-il de paramètres libres?

   (b) Si N est très grand, comment pourriez vous augmenter la capacité?

(c) Si $N = 200$, pensez-vous que la capacité sera trop grande, adéquate, ou trop petite?

(d) Supposons que vous mesuriez après apprentissage une très faible erreur sur les 150 exemples d'apprentissage mais une erreur beaucoup plus grande sur les 50 exemples restant (ensemble de validation). Suggérez un ensemble d'expériences différentes (i.e., correspondant à differents ensembles de solutions) que vous pourriez faire pour essayer d'améliorer la généralisation.

(e) Si vous avez essayé beaucoup de solutions différentes en (d), et que vous avez choisi la solution qui fonctionnait le mieux sur les 50 exemples de validation, que pouvez-vous dire de l'erreur obtenue sur ces 50 exemples avec cette solution?