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 H96**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A94: 20 Mars 1996 HEURE: 8h30 - 10h20
SALLE: P312

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

1. Expliquez brièvement et qualitativement pourquoi *généraliser* est intrinsèquement difficile.

2. Que peut-on dire sur l'ensemble des fonctions que peut réaliser

   (a) un neurone formel (sortie discrète)?

   (b) plusieurs neurones formels (sortie discrète) avec les mêmes entrées (non-connectés entre eux)?

   (c) un réseau de neurones formels (sortie discrète) avec au moins deux couches cachées?

   (d) un réseau de neurones multi-couche (activation sigmoide) avec au moins une couche cachée?

3. Considérez les facteurs qui suivent; pour chacun d'eux, dites si son effet serait plutôt de *réduire*, ne *pas changer*, ou bien *augmenter* la **capacité effective** d'une machine à apprendre entrainée par descente de gradient stochastique. Expliquez brievement votre réponse.

   (a) Augmenter le nombre de paramètres libres.

   (b) Lisser l'entrée (e.g., une image ou une séquence).

(c) Faire la moyenne des sorties (continues) de cette machine, déjà entrainée, avec les sorties (aussi continues) d'une autre machine, entrainée sur un ensemble d'entrainement différent (pour la même tâche).

(d) Utiliser un algorithme d'optimisation du critère de coût qui s'approche plus vite et plus précisement d'un minimum du coût sur les examples d'entrainement. Supposez que ce nouvel algorithme tombe dans le même bassin minimum que la descente de gradient stochastique mais qu'il converge précisement au minimum alors votre descente de gradient stochastique n'y parvient pas (à cause du "bruit" inhérent à la descente de gradient stochastique).

4. Au fur et à mesure que l'on augmente le nombre d'exemples d'apprentissage disponibles, que peut-on dire de la capacité optimale?

5. Quelle est la différence entre gradient stochastique et gradient *batch* (ou gradient complet)? Quels sont les avantages du gradient stochastique?

6. Expliquez pourquoi selon vous l'algorithme du perceptron est plus **discriminant** qu'une régression

linéaire (avec sortie désirées -1 et 1 pour les deux classes, par exemple).

7. Essayez d'imaginer une manière de généraliser l'algorithme des **K-plus-proches voisins** (un algorithme de classification) à des problèmes de prédiction (régression non-linéaire). Ecrivez formellement les équations de la valeur prédite pour votre algorithme et expliquez comment l'utiliser.

8. Qu'arrive-t-il à la fonction calculée par un réseau multicouche quand ses poids sont tous près de zero?

9. Qu'arrive-t-il à la capacité et à la fonction calculée par un réseau multicouche quand ses poids deviennent très grands?

10. Qu'est-ce que le partage de poids?

11. Un **réseau récurrent** peut être vu comme un réseau multi-couche où chaque couche est associée à un temps t, et où les poids sont partagés entre les differents t. Il s'agit en fait d'un réseau à délais avec des cycles. Supposez le cas le plus simple, c'est à dire avec seulement deux neurones: un neurone d'entrée, et un neurone caché avec une connection à délai 0 provenant du neurone d'entrée, et une connection à délai 1 provenant de lui-même. Dessinez le réseau déplié dans le temps, montrant ainsi les activations de chaque neurone aux temps $t = 0, 1, 2, ...$, ce qui est un graphe d'influence de ce réseau. Commentez sur la manière de calculer le gradient sur les 3 paramètres libres (2 poids et un biais, partagés dans le temps).