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 A04**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN FINAL A04: 7 décembre 2004 HEURE: 13h - 15h
SALLE: PAA-3195

DIRECTIVES PEDAGOGIQUES:
- Une feuille recto-verso permise
- Répondre directement sur le questionnaire.
- 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.

1. **Validation croisée et sélection de modèles**

   - Plusieurs algorithmes d'apprentissage itératifs peuvent être améliorés en effectuant un "arrêt prématuré", c'est à dire en gardant le modèle $A_{t^*}(D)$ correspondant à l'itération $t = t^*$ qui a donné la meilleure erreur $e_{t^*} = \min_t e(A_t(D), D_{valid})$ sur un ensemble de validation D_{valid} indépendant de l'ensemble d'apprentissage D (toutes les données sont supposées i.i.d.). Supposons qu'on estime *l'erreur de généralisation de* $A_{t^*}(D)$ (en espérance sur D, D_{valid} et des exemples de test) par l'erreur de validation e_{t^*} mesurée à $t = t^*$. Le biais de cet estimateur est-il optimiste, neutre ou pessimiste? Pourquoi?

- L'algorithme de sélection de modèle A est défini en fonction de sous-modèles A_1, ..., A_n: il choisit un des A_i par validation croisée à k partitions (*k-fold cross-validation*). Pour estimer l'erreur de généralisation de A, on peut donc effectuer de la validation croisée **double** (avec encore k partitions au niveau supérieur). Montrez que cette méthode fournit un estimateur non-biaisé de l'erreur de généralisation de A (appliqué à un ensemble de données d'une certaine taille, tirées *i.i.d.* d'une loi P).

- Dans la question précédente, pourquoi ne peut-on pas se contenter d'une validation croisée ordinaire, c'est à dire d'estimer l'erreur de généralisation de A par le minimum sur i de l'erreur de validation croisée de A_i sur D? Dans cette réponse nous ne nous intéressons pas au biais dû à la petite variation de taille de l'ensemble d'apprentissage (pour A vs un A_i). Une réponse intuitive suffira.

2. **Régularisation et inférence Bayesienne**.

- Soit le critère d'apprentissage régularisé suivant:

$$f_a^* = \text{argmin}_f \left(\lambda \Omega(f) + \frac{1}{n} \sum_i L(f, z_i) \right)$$

où f représente une fonction solution possible, $D = \{z_i\}_{1 \leq i \leq n}$ l'ensemble d'apprentissage d'exemples i.i.d., $L(f, z)$ la perte encourue quand z est réalisé alors qu'on utilise la solution f, λ le coefficient de régularisation, et $\Omega(\cdot)$ une fonctionnelle qui permet d'établir une préférence entre différentes solutions. Montrez le détail de l'équivalence qui existe entre un tel critère de régularisation et la solution qui serait choisie (si on devait en choisir une en particulier) selon l'approche Bayesienne, c'est à dire celle qui est la plus probable étant donné les données: $f_b^* = \text{argmax}_f P(f|D)$.

- Avec le critère ci-dessus, on doit choisir l'hyper-paramètre λ. Supposons que $\lambda = \lambda_1$ soit optimal (dans le sens de minimiser l'erreur de généralisation) pour D contenant n exemples d'apprentissage. Si on multiplie n par 10, est-ce que le λ optimal λ_2 sera inférieur, égal, ou supérieur à λ_1? Pourquoi? (un argument intuitif mais correct suffira).

3. **Truc du noyau et ACP**. Dans l'analyse en composantes principales (ACP) à noyau on effectue l'ACP dans l'espace des $\phi(x)$, avec un noyau $K(x, y) = \phi(x) \cdot \phi(y)$. Mais pour ce faire il faut transformer le noyau K en noyau dépendant des données $\tilde{K}(x, y) = \tilde{\phi}(x) \cdot \tilde{\phi}(y)$, où $\tilde{\phi}$ a la propriété d'être centré par rapport aux données, i.e., $\sum_{i=1}^{n} \tilde{\phi}(x_i) = 0$, avec $D = \{x_1, \ldots, x_n\}$ l'ensemble d'apprentissage. Les $\tilde{\phi}(x)$ sont reliés aux $\phi(x)$ par simple translation. Montrez comment obtenir ce noyau centré $\tilde{K}$ par une simple opération sur K avec les données, sans jamais avoir besoin de calculer des $\phi(x)$ explicitement. Il ne suffit pas de donner la formule (qui était dans les notes) mais il faut aussi prouver qu'elle donne la bonne propriété.

   QUESTION BONUS: Que se passe-t-il si on ne centre pas le noyau?

4. **Malédiction de la dimensionalité**. On peut montrer qu'un algorithme d'apprentissage de variété comme LLE, Isomap ou l'ACP à noyau (avec noyau gaussien) a la propriété suivante: il approxime localement la forme de la variété autour du point x par un plan dont les vecteurs directeurs (donc les vecteurs tangents à la variété au point x) sont des combinaisons linéaires des vecteurs de différence $x - x_i$, avec les x_i des exemples d'apprentissage voisins de x. Soit $x \in \mathbf{R}^m$ et supposons que la variété soit de dimension $d < m$. Pour approximer raisonnablement la variété autour de x, il faut donc au moins d voisins de x dans l'ensemble d'apprentissage, assez près de x pour que l'approximation de la variété par un plan soit assez bonne dans la boule autour de x qui contient les voisins x_i. Étant donné que la variété n'est pas globalement plane mais seulement localement linéaire, il faut que ces boules soient assez petites, selon la courbure de la variété. Soit r le rayon de ces boules qui garantit que l'approximation de linéarité locale est acceptable. Que peut-on dire sur l'effet de la malédiction de la dimensionalité en ce qui concerne ces algorithmes? Plus précisément, comment est-ce que la qualité de l'approximation ou le nombre d'exemples nécessaires évoluent quand m **ET/OU** d augmentent, et pourquoi? Pour simplifier l'argumentation on peut supposer une distribution uniforme sur la variété.

   QUESTION BONUS: si les points ne sont pas exactement sur la variété mais subissent un bruit uniforme ϵ qui affecte toutes les m dimensions, près de la variété, comment pensez-vous que cela influence le résultat?

5. **Data-mining**. Dans le processus méthodologique de data-mining nous avons vu qu'il était important de détecter les variations de distribution des données dans le temps (par rapport à la date associée à chaque donnée).

   - Pourquoi pensez-vous que cela puisse être utile et comment pourrions-nous utiliser cette information pour obtenir de meilleures prédictions? **Considérez deux cas de figure**: un changement *brusque* de distribution dans le temps, ou une variation lente et *graduelle* due à une tendance générale (comme l'inflation).

- Essayez d'imaginer (et expliquez) un algorithme permettant de vérifier s'il existe une variation de la distribution en fonction du temps.