Prénom et nom de l'étudiant: _____

Code permanent: _____

IFT 3395 ou 6390?

---

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

TITRE DU COURS: **Fondements de l'apprentissage machine**
SIGLE DU COURS : **IFT3395/6390 A10**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A10: 20 octobre 2010 HEURE: 14h30 - 16h15
SALLE: AA-1175

DIRECTIVES PÉDAGOGIQUES: - Documentation permise.
- Répondre directement sur le questionnaire. Vous pouvez utiliser l'arrière des pages aussi si vous en avez besoin.
- 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.**
- Échanger des informations lors d'un examen (ou autres formes de tricherie) est du **plagiat**, qui est passible de sanctions allant jusqu'à l'exclusion du programme.
- Suggestion: commencez par tout lire, et faire les questions les plus faciles.

---

1. (9 points) Nommez et écrivez une formule pour des fonctions de perte, une pour la classification, une pour la régression, et une pour l'estimation de densité. Si vous en trouvez plus, expliquez-les et gagnez des points boni.

2. (6 points) En utilisant **une** des formules ci-haut, écrivez le risque empirique moyen correspondant à cette fonction de perte.

3. (5 points) Pour cette même fonction de perte, écrivez l'erreur de généralisation correspondante.

4. (10 points) Que pouvez-vous dire sur le lien et la différence qui existe entre ces deux expressions (le risque empirique moyen et l'erreur de généralisation)? Considérez en particulier deux cas de figure: (1) les données utilisées sont choisies indépendamment du prédicteur (comme une erreur de test), et (2) on a minimisé ce risque empirique en ajustant des paramètres du prédicteur (comme une erreur d'apprentissage). **BONUS**: qu'arrive-t-il quand le nombre d'exemples tend vers l'infini?

5. (5 points) Qu'arrive-t-il en général à l'erreur d'apprentissage quand on augmente le $k$ du classifieur par $k$-plus-proches voisins?

6. (5 points) À quelle genre de relation s'attendre entre l'erreur de généralisation et $k$ du $k$-plus-proches voisin? Vous pouvez répondre à la question en dessinant une courbe (nommez les axes!).

7. (7 points) Si le vecteur d'entrée $X$ est dans $\mathbf{R}^d$, quel est le nombre de paramètres à ajuster pour un classifieur linéaire appliqué à $X$? à quoi correspondent-ils, géométriquement?

8. (5 points) Dessinez un ensemble de données en 2 dimensions, pour un problème de classification, avec la propriété que cet ensemble n'est pas linéairement séparable.

9. (20 points) Écrivez le pseudo-code pour construire un classifieur de Bayes pour 2 classes, en utilisant comme estimateur de densité les fenêtre de Parzen, avec un $\sigma$ que vous devez choisir en utilisant les données, en séparant les données fournies en partie apprentissage et partie 'validation'. La fonction `train` prends donc une matrice `X` (une rangée par exemple) et un vecteur `Y` (un élément 1 ou -1 par exemple). Votre pseudo-code doit essayer les valeurs de $\sigma$ suivantes: 1,.1,.01,.001, et choisir celle qui donne la meilleure log-vraisemblance de validation.

10. (5 points) Si on peut minimiser un critère d'apprentissage $C$ (risque empirique) analytiquement par rapport à un vecteur de paramètres $\theta$, quelle **équation** doit-on résoudre? **BONUS**: Quelle **inégalité** doit-on vérifier?

11. (5 points) En général, à quoi s'attendre concernant l'erreur de généralisation d'un algorithme d'apprentissage au fur et à mesure qu'on augmente le nombre d'exemples d'apprentissage? Augmente, reste la même ou diminue?

12. (8 points) **(a)** Quel est la complexité du temps de calcul (en notation $O(\cdot)$) nécessaire à l'entraînement d'un modèle de régression linéaire, quand il y a $n$ exemples d'entraînement et $d$ dimensions d'entrée (et une seule dimension en cible)?

Notez qu'une inversion de matrice $N \times N$ se fait en temps $O(N^3)$. **(b)** Quelle serait la complexité de calcul si on fait de la validation croisée à $K$ plis pour estimer l'erreur de généralisation?

13. (10 points) Supposons que nous avons un algorithme d'apprentissage pour la classification binaire qui est basé sur l'estimation de $f(x) = P(Y = 1|X = x)$, où $Y \in \{0, 1\}$. Considérez comment convertir $f(x)$ en décision binaire, et prouvez que $1_{f(x)>0.5}$ est un choix qui minimise l'espérance de l'erreur de classification si on prend pour vrai l'estimateur de $P(Y|X)$ produit par l'algorithme d'apprentissage.

14. (15 points, **BONUS**) Vous allez généraliser l'algorithme de régression par fenêtres de Parzen, qui estime $E[Y|X]$ (pour $Y$ scalaire ou vecteur), à un algorithme qui estime la **matrice** de covariance conditionnelle des $Y$ (qui est donc un vecteur), étant donné $X$. L'algorithme estime donc $E[(Y - E[Y|X])(Y - E[Y|X])'|X]$, soit la matrice dont les éléments sont $E[(Y_i - E[Y_i|X])(Y_j - E[Y_j|X])'|X]$. Ici on dénote $v'$ la transposée de $v$ (donc un vecteur rangée si $v$ est colonne) et $vv'$ la matrice dont l'élément $(i, j)$ est $v_i v_j$. La fonction de $X$ apprise et produite est donc une matrice

avec un élément pour chaque paire d'éléments de $Y$. Expliquez le principe de fonctionnement de votre algorithme (formules, pseudo-code, comme vous voulez). Notez que c'est un problème intéressant en finance computationnelle.