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 A03**

NOM DU PROFESSEUR: Yoshua Bengio

DATE DE L'EXAMEN INTRA A03: 27 octobre 2003 HEURE: 15h30 - 17h25
SALLE: Z-337

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. Vous allez généraliser l'algorithme des K-plus-proches voisins à un problème de décision financière en gestion de portefeuille. Les données D dans ce problème sont des paires de vecteurs, $D = \{(x_t, r_t)\}$, où x_t est un vecteur d'entrée dans $\mathbf{R}^p$ et r_t un vecteur dans $\mathbf{R}^m$ de rendements observés sur m actifs. L'estimateur recherché $\hat{y}(x)$ devrait être tel qu'il maximise le profit et qu'il minimise la variance du profit sur les k voisins de x. Plus précisément, soit $p_t = w.r_t$ le profit sur le voisin t si on choisit de placer nos actifs selon le vecteur de pondérations w, soit $\bar{p}(x)$ la moyenne de p_t sur les k voisins de x et $s^2(x)$ sa variance empirique (estimée sur les vecteurs r_t des mêmes k voisins). Alors le critère à minimiser localement (pour ce x particulier) est $C_x(w) = -\bar{p}(x) + \alpha s^2(x) + \lambda ||w||^2$ (par opposition à l'erreur quadratique sur les voisins de x, avec le K-plus-proches voisins classique) où α et λ sont des scalaires donnés. Le dernier terme est un terme de régularisation. Notez bien qu'on peut écrire $C_x(w)$ de manière vectorielle, ce qui simplifie beaucoup le calcul: $\bar{p} = -w.\bar{r}$ où $\bar{r}$ est la moyenne empirique des vecteurs de rendements r_t des k voisins, et $s^2 = w'Aw$ où A est la matrice de covariance empirique des mêmes vecteurs.

   (a) Écrivez une formule analytique qui donne l'estimateur $\hat{y}(x)$ (le w qui minimise $C(w)$) en fonction de l'ensemble de données D.

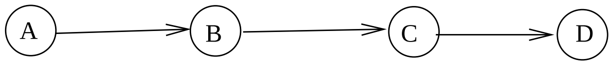
(b) Dans la solution ci-haut, considérez ce qui arriverait si on n'avait pas mis de terme de régularisation et que le nombre de voisins k était plus petit que le nombre d'actifs m. Même si $k > m$, choisir $\lambda > 0$ serait probablement utile. Pourquoi?

(c) Dans ce modèle, il y a plusieurs hyper-paramètres (variables à choisir avant de pouvoir appliquer l'algorithme): λ, α, k. Décrivez qualitativement l'effet de varier chacun.

2. Si on observe qu'en augmentant un peu un hyper-paramètre qui influence la capacité, l'erreur d'apprentissage diminue mais l'erreur de généralisation estimée augmente, peut-on dire qu'on est en plutôt zone d'*overfitting* (sur-généralisation) ou en zone d'*underfitting* (sous-généralisation)? En conséquence devrait-on essayer d'augmenter ou de diminuer la capacité?

3. En variant un hyper-paramètre on observe que l'erreur d'apprentissage diminue. Quel est l'effet probable sur le biais et la variance de l'estimateur?

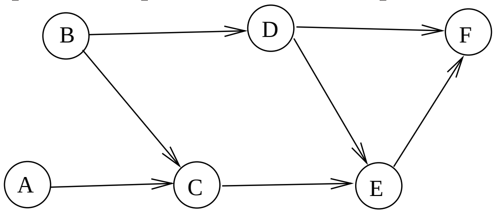
4. Considérez le modèle graphique ci-bas. Les paramètres du modèle sont les tables de probabilités $P(A = a)$, $P(B = b|A = a)$, $P(C = c|B = b)$, $P(D = d|C = c)$. Soit n_a, n_b, n_c, n_d respectivement le nombre de valeurs que peuvent prendre A, B, C, et D.



   (a) Comment exprime-t-on la distribution jointe en terme de ces tables?

(b) Supposons qu'on observe la variable D, et qu'on veuille calculer la vraisemblance de l'évènement $P(D = d)$. Si on écrit brutalement la marginalisation de la probabilité jointe, on a une somme avec $n_a \times n_b \times n_c$ termes. Montrer comment on pourrait calculer $P(D = d)$ efficacement, c'est à dire en exploitant les indépendences conditionnelles, grâce à une factorisation de cette somme. Ce calcul efficace devrait pouvoir se faire en effectuant $O(n_c + n_b \times n_c + n_a \times n_b)$ opérations. Il est important de bien spéficier l'ordre dans lequel faire les calculs pour arriver à un algorithme efficace.

5. Dans le modèle graphique ci-bas, dites pour chaque énoncé d'indépendance conditionnelle ci-bas si il est vrai ou faux:



   (a) $B \perp E | C, D$?
   (b) $C \perp D | A, B, E, F$?
   (c) $C \perp D | B$?
   (d) $B \perp F | D$?
   (e) $A \perp B$?
   (f) $A \perp B | C$?

6. Dans certaines applications de la classification binaire, on ne connaît pas la classe désirée pour chaque exemple mais seulement pour chaque groupe d'exemples: les exemples sont organisés en groupes, et on sait pour chaque groupe si **au moins un** des exemples est de la classe 1 (alors on étiquette le groupe par un $y = 1$, sinon on étiquette le groupe par $y = 0$). Les données sont donc un ensemble de groupes $\{g_1, \ldots, g_n\}$, avec pour chaque groupe $g_i = (y_i, \{x_{i1}, x_{i2}, \ldots, x_{im}\})$ une étiquette de groupe y_i (0 ou 1) et un ensemble de vecteurs d'entrées (x_{ij} est le j-ème exemple du groupe i). Inventez pour ce type de problème un nouveau critère d'apprentissage (qu'on pourrait par exemple utiliser pour entraîner un réseau de neurones), pour remplacer le critère classique pour entraîner un réseau de neurone en classification probabiliste ($-\sum_i \log P(\text{classe}_i|\text{entree}_i)$). Cette question demande surtout de la logique et de l'imagination. Justifiez votre choix.

7. Considérez la classe suivante de fonctions de $\mathbf{R}$ vers $\mathbf{R}$:

$$f_\theta(x) = \sum_{i=1}^{m} e^{a_i} g(b_i + e^{c_i} x)$$

avec paramètres $\theta = (a_1, \ldots, a_m, b_1 \ldots b_m, c_1 \ldots c_m)$ et $g(z) = 1/(1 + e^{-z})$ la fonction sigmoide. L'entier m est un hyper-paramètre (comme le nombre d'unités cachées d'un réseau de neurones). Pour simplifier la correction, dénotons $s_i = b_i + e^{c_i} x$ et $t_i = g(s_i)$.

(a) Dessinez un graphe de flot avec un noeud pour chaque scalaire ci-haut impliqué dans le calcul de $E = (f_\theta(x) - y)^2$ (avec des ... pour indiquer l'énumeration des noeuds indicés de 1 à m). Proposez un ordre de visite des noeuds (les a_i, b_i, c_i, s_i, t_i en plus de x, $f(x)$ et l'erreur E) permettant de calculer E étant donné les θ et x.

(b) Montrez comment calculer $\frac{\partial E}{\partial v}$ pour chaque noeud v du graphe de flot, en traversant le graphe dans l'ordre inverse de celui défini ci-haut. Utilisez la recette décrite en classe, c'est à dire qu'à chaque noeud on a une structure de données avec un champs v et un champs $\frac{\partial E}{\partial v}$, ainsi qu'une procédure qui calcule $\frac{\partial E}{\partial v}$ étant données les valeurs de $\frac{\partial E}{\partial u_i}$ pour les successeurs u_i de v dans le graphe.

(c) Cette classe de fonctions a plusieurs propriétés intéressantes. Premièrement, quel que soit le choix des paramètres, la relation entre x et $f_\theta(x)$ satisfait une contrainte de dérivée. Laquelle?

(d) (QUESTION BONI). Cette classe de fonctions a une autre propriété intéressante: pour tout ϵ et tout intervalle I, il existe m et θ tel qu'on puisse approximer une fonction donnée $h(x)$ (continue et satisfaisant la contrainte de dérivée ci-haut) avec une erreur $|h(x) - f_\theta(x)|$ plus petite qu'ϵ pour x dans l'intervalle I. De quelle manière vous y prendriez vous pour prouver ce théorème? Suggestion: faites un dessin en escalier dont chaque marche est créée par une sigmoide au bon endroit et à la bonne hauteur.