Département d'informatique et de recherche opérationnelle Automne 2006

IFT 6266
Algorithmes d'apprentissage

Yoshua Bengio

Bureau : PAA #3339, courriel: pift6266@iro

Devoir #4

Donné le 16 novembre 2006, Dû le 7 décembre 2006

1. Considérez un algorithme à noyau (i.e., $f(x) = \sum_i \alpha_i K(x, x_i)$) dans lequel on a la contrainte $\sum_i \alpha_i = 0$ (c'est le cas des SVMs). Pour le noyau Gaussien $K(u, v) = e^{-||u-v||^2/\sigma^2}$, prouvez que le classifieur binaire correspondant (qui prend le signe de $f(x)$) converge vers un classifieur linéaire (et pas seulement **constant**) quand σ devient de plus en plus grand.

2. Vous allez appliquer le truc du noyau à la régression logistique: $f(x) = sigmoid(w'x)$ (ignorez le biais). Les données d'apprentissage sont $D = \{(x_i, y_i)\}_{i=1}^n$, avec $y_i \in \{0, 1\}$.

   (a) Montrez que la solution de la régression logistique (après convergence de l'optimisation) peut s'écrire comme une combinaison linéaire des x_i d'apprentissage (avec des poids qui pourraient dépendre de tous les x_i de façon non-linéaire).

   (b) Utilisez ce fait pour écrire $w = \sum_i \alpha_i x_i$. Changez la formulation en remplaçant les x et les x_i par des $\phi(x)$ et $\phi(x_i)$, avec la définition $K(u, v) = \phi(u)'\phi(v)$. Montrez que l'on a maintenant un problème d'optimisation convexe en α.

3. Dans le modèle probabiliste graphique ci-dessous vous allez répondre à des questions concernant l'indépendence conditionnelle entre différents groupes de variables aléatoires. Le modèle concerne les variables aléatoires A, B, C, D, E, F et G. Le graphe contient les arcs suivants: $A \to B$, $A \to C$, $A \to E$, $B \to D$, $C \to E$, $D \to E$, $D \to G$, $E \to F$, $F \to G$. Dessinez le graphe. Expliquez pourquoi chacun des énoncés ci-bas est vrai ou faux. Il suffit de mentionner une condition SUFFISANTE pour que l'énoncé soit vrai ou faux.

   (a) $C \perp D$

   (b) $C \perp D|A$

   (c) $C \perp D|B$

   (d) $C \perp D|A, G$

   (e) $C \perp D|E, B$

   (f) $A \perp G|D, F$

(g) $A \perp G | E, B$

4. Vous allez appliquer l'algorithme EM à un mélange de lois discrètes. La variable observée est le vecteur binaire X (i.e. $x_i \in \{0, 1\}$, $i \in \{1, \ldots, d\}$) et la variable cachée est la variable discrète Z, prenant ses valeurs dans $\{1, \ldots, N\}$. La loi marginale de Z est multinomiale et $P(X|Z)$ a la forme particulière suivante:

$$P(X = (x_1, \ldots, x_d)|Z = z) = P(x_1|z)P(x_2|x_1, z)P(x_3|x_2, z)P(x_4|x_3, z) \ldots P(x_d|x_{d-1}, z)$$

où l'on paramétrise toutes ces probabilités par des tables. Les paramètres libres sont donc les $P(Z = i)$, les $P(X_1 = 1|Z = j)$, et tous les $P(X_r = 1|X_{r-1} = j, Z = k)$, avec la contrainte habituelle de somme à 1 pour les $P(Z = i)$.

   (a) Montrez comment calculer les postérieurs $P(Z = i|X)$.

   (b) Écrivez la fonction auxiliaire, en y ajoutant un terme $\lambda((\sum_i P(Z = i)) - 1)$ pour la contrainte.

   (c) Trouvez la formule de ré-estimation des paramètres en maximisant analytiquement cette fonction auxiliaire augmentée.

5. (**BONUS**) Ce problème est lié à la compréhension de l'analyse en composantes principales. Considérez le réseau de neurones auto-associatif à deux couches LINÉAIRES, avec moins d'unités cachées h que d'entrées d, et les contraintes que la matrice de la deuxième couche est la transposée de celle de la première couche et que les rangées de cette matrice sont orthonormales. Plus formellement, la sortie est

$$f(x) = W'Wx$$

avec W de dimensions $h \times d$, et $WW' = I$. Supposons aussi que les entrées ont une moyenne 0 (on a préalablement soustrait leur moyenne). Montrez alors que si on minimize l'erreur quadratique de reconstruction

$$\sum_i ||f(x_i) - x_i||^2$$

sous la contrainte $WW' = I$, alors on obtient que les rangées de W sont les vecteurs propres principaux de la matrice de covariance empirique des données $\{x_1, \ldots x_n\}$.
Indice: trace$(ABCD)$ **=trace**$(CDAB)$ **et scalaire=trace(scalaire).**