Département d'informatique et de recherche opérationnelle Automne 2005

IFT 6266
Algorithmes apprentissage

Yoshua Bengio

Bureau : Math-Info #3339, e-mail: bengioy@iro, #tel: 343-6804

Devoir #1

Donné le 7 septembre 2006, Dû le 21 septembre 2006

1. Expliquez pourquoi selon vous **généraliser** correctement uniquement à partir d'un ensemble d'exemples est fondamentalement un problème qui est mathématiquement mal défini.

2. Supposons que le seul algorithme d'apprentissage que vous connaissiez soit la régression linéaire, qui estime $E[Y|X]$ à partir de paires (x_i, y_i). Si on vous demande de vous attaquer à un problème de classification probabiliste, i.e. l'estimation de $P(Y|X)$, avec les $y_i \in \{1, \ldots, N\}$, comment est-ce que vous pourriez vous y prendre en utilisant seulement votre algorithme de régression linéaire? Commencez par considérer le cas plus simple de $N = 2$ classes. Selon vous, quelles sont les défauts de cette approche?

3. Considérons un classifieur affine, i.e. $f(x) = \text{sign}(b + w'x)$, avec paramètres b (un scalaire) et w (un vecteur), et entrée x. Montrez que $\frac{|b+w'x|}{||w||}$ est la distance entre x et la surface de décision $b + w'x = 0$, et que le signe $(f(x))$ indique de quel côté de la surface x se trouve.

4. Vous allez ici généraliser la régresssion linéaire ordinaire dans deux directions. Premiérement vous allez considérer la possibilité que l'erreur sur certains exemple soit plus importante que l'erreur sur d'autre. On va donc supposer qu'on nous donne un poids v_t sur l'erreur de l'exemple t. Par ailleurs on va supposer que certaines solutions (w, b) sont préférables à d'autres (technique de la régularisation) et on va pénaliser les moins préférables. Plus précisement vous allez considérer que l'on veut minimiser λ fois la norme au carré de w ($\lambda \sum_i w_i^2$) en plus des erreurs de prédiction. Le coût d'apprentissage est donc $C = \lambda \sum_i w_i^2 + \sum_t v_t(b + w'x_t - y_t)^2$. Dérivez la solution analytique à ce problème de minimisation (et donc donnez une formule pour w et b qui minimise C).

5. Un algorithme d'optimisation est dit invariant linéaire si une transformation linéaire de la paramétrisation ne change pas la solution. Plus précisément, soit $C(\theta)$ le coût à

minimiser en θ, et soit $\tilde{\theta} = A\theta$ avec une matrice A inversible. Soit θ^* la solution obtenue par un algorithme d'optimisation qui cherche à minimiser $C(\theta)$ en θ. Si l'algorithme est appliqué à minimiser $C(A^{-1}\tilde{\theta})$ en $\tilde{\theta}$ et qu'il obtient $\tilde{\theta}^* = A\theta^*$, on dit que l'algorithme est invariant linéaire. Dans le cas d'un algorithme itératif, cela doit être vrai après chaque itération de l'algorithme. Chaque itération de la descente de gradient de $C(\theta)$ sur le vecteur $\theta \in \mathbb{R}^n$ modifie θ ainsi:

$$\theta \leftarrow \theta - \epsilon \frac{\partial C}{\partial \theta}$$

avec ϵ une petite constante positive. Chaque itération de la méthode de Newton modifie θ ainsi:

$$\theta \leftarrow \theta - (\frac{\partial^2 C}{\partial \theta^2})^{-1} \frac{\partial C}{\partial \theta}$$

Est-ce que la descente de gradient est invariante linéaire? Est-ce que la méthode de Newton est invariante linéaire? Donnez le développement. Si vous n'y arrivez pas avec la méthode de Newton, considérez au moins le cas où θ est un scalaire.

6. Considérons le problème de prédiction non-paramétrique dans des espaces discrets. Soit la variable d'entrée $X = (X_1, \ldots, X_d)$ avec $X_i \in \{1, \ldots, V\}$ et la variable de sortie $Y \in \{0, 1\}$. On estime $E[Y|X = x_{n+1}]$ par la moyenne des y_t observés dans le sous-ensemble $\mathcal{V}(x_{n+1})$ des n exemples d'apprentissage (x_t, y_t) pour lesquels $\sum_i 1_{x_{t,i} \neq x_{n+1,i}} < \epsilon$. Si on suppose les X uniformément distribués, on voudrait pouvoir dire quelque chose d'intéressant sur la probabilité que $\mathcal{V}(x_{n+1})$ soit vide. On s'intéresse à ce qui se passe quand d devient grand.

   (a) Considérez d'abord le cas plus simple $\epsilon = 1$ (égalité) et $V = 2$ (symboles binaires). (Si vous avez de la difficulté avec les maths, essayer d'y répondre de manière expérimentale comme en **(c)**).

   (b) **BONUS**. Trouvez une formule pour le cas plus général $\epsilon > 1$ (match inexact).

   (c) **BONUS**. Étudiez empiriquement le problème en faisant un histogramme de cette probabilité en fonction de ϵ, pour plusieurs valeurs de d de plus en plus grandes. Vous pouvez soit le faire avec la formule trouvée en **(b)** ou bien en faisant des tirages aléatoires des X_i et du point test X.