Examen Intra IFT3395/6390

Mardi 19 février 2008

Durée: 1h45

Professeur: Pascal Vincent

Prénom:

Nom:

Code permanent:

IFT 3395 ou 6390?

L'examen est long, alors soyez bref mais précis.
Veuillez répondre directement dans les blancs laissés à cet effet.
Bonne chance!

Notations

Les notations suivantes sont définies pour tout l'examen:
On suppose qu'on dispose d'un ensemble de données de n exemples: $D_n = \{z_1, ..., z_n\}$. Dans le cas supervisé chaque exemple z_i est constitué d'une paire *observation, cible:* $z_i = (x_i, t_i)$, alors que dans le cas non-supervisé, on n'a pas de notion de cible explicite donc juste un vecteur d'observation: $z_i = x_i$. On suppose que chaque observation est constituée de d traits caractéristiques: $x_i \in \mathbb{R}^d$.

1 Sélection de modèle (15 pts)

Vous êtes embauché dans une entreprise qui réalise des systèmes de vérification d'identité et qui travaille sur un nouveau système de reconnaissance de visages pour un important client. Le système doit être capable de distinguer une dizaine de personnes autorisées et les différencier de toute autre personne non autorisée. L'entreprise dispose d'une base de donnée comportant 200 000 images de visages étiquetés (identifiés comme autorisé ou non autorisé). Un collègue vient vous voir et vous dit qu'il a essayé 3 variantes d'algorithme de classification (par exemple des réseaux de neurones avec un nombre différent de neurones cachés), qu'il a entraîné sur ces 200 000 images. Le premier obtenait 4% d'erreur, le second 2%, et le 3ème 0.3% d'erreur sur les 200 000 images. Puisque son expérience montre clairement que le 3ème a une performance bien meilleure, c'est donc celui-là qu'il veut utiliser dans le nouveau système.

1.1 (5 pts) Êtes-vous d'accord avec lui? Expliquez/justifiez votre réponse.

1.2 (5 pts) Si vous n'êtes pas d'accord, comment proposeriez-vous à votre collègue de procéder pour décider laquelle des variantes utiliser? Expliquez en détail.

1.3 (5 pts) Le client vous demande une estimation fiable de la performance à laquelle il pourra s'attendre du système sur le terrain. Comment vous y prendriez-vous pour la lui fournir?

2 Classification, régression, estimation de densité (20 pts)

2.1 (10 pts) Dans le cadre général donné ci-dessus, expliquez brièvement la différence entre un problème de classification binaire, un problème de classification à m classes, un problème de régression, et un problème d'estimation de densité de probabilité. Indiquez spécifiquement pour *chacun* de ces types de problèmes: la nature de la cible t_i (quelle plage de valeurs elle peut prendre); la nature de la sortie que l'algorithme va calculer pour pour un point de test x; la formule du *risque empirique* (le coût moyen sur l'ensemble d'apprentissage) que l'algorithme d'apprentissage devrait idéalement optimiser, incluant une fonction de coût ou de perte précise (différente pour chaque cas).

2.2 (10 pts) On suppose qu'on a un problème de classification à m classes. Pour ce problème, on a entraîné m estimateur de densité $f_k(x)$, $k \in \{1, ..., m\}$. Chaque f_k a été entraîné sur le sous ensemble des entrées de D_n de classe k (ce sous-ensemble étant de taille n_k). Expliquez en détail, et en équations, comment on peut se servir des f_k pour classifier des nouveaux points de test x (sans oublier de préciser comment vous estimeriez tout paramètre additionnel nécessaire). Comment appelle-t-on ce genre de classifieur?

3 K plus proches voisins (20 pts)

Dans ce qui suit, on supposons qu'on a un problème avec $m = 10$ classes, en dimension d=200, et que notre ensemble de d'entraînement D_n comporte exactement 2000 exemples de *chaque* classe. On suppose de plus que tous ces exemples sont différents. On s'intéresse à l'algorithme des K plus proches voisins (K-PPV ou K-NN) avec une distance Euclidienne.

3.1 (4 pts) Expliquez brièvement mais clairement l'algorithme de classification des K plus proches voisins (K-NN) pour un tel problème. Expliquez précisément dans vos propres mots comment l'algorithme prend sa décision pour un point de test x.

3.2 (4 pts) Quelle est la complexité algorithmique du calcul de la classification d'un nouveau point x (soyez le plus précis possible en fonction des paramètres du problème)?

3.3 (4 pts) Indiquez la plus petite et la plus grande valeur possible de K qu'on peut logiquement utiliser pour ce problème. Dans chaque cas donnez le taux d'erreur de classification (en %) qu'on s'attend à obtenir sur l'ensemble d'apprentissage.

3.4 (4 pts) Tracez sur un graphique l'allure typique des courbes d'erreur d'apprentissage (en trait plein) et de test (en trait pointillé). Prenez soin d'indiquer précisément sur chaque axe la quantité qu'il représente, de graduer les axes, et de placer clairement les points dont on peut connaître la valeur exacte pour notre problème.

3.5 (4 pts) L'algorithme K-NN est considéré comme une méthode de type "non-paramétrique". Indiquez deux autres méthodes considérées non-paramétriques (on ne demande pas d'expliquer leur fonctionnement).

4 Frontière de décision et séparabilité linéaire (20 pts)

4.1 (4 pts) Donnez l'équation d'une fonction discriminante linéaire. A quoi une telle fonction peut-elle servir? Précisez quels sont ses paramètres et quelle est leur dimension (pour chacun).

4.2 (3 pts) Expliquez en français, précisément, mais dans vos propres mots la notion de région de décision. Il s'agit d'une région de quel espace? Combien y a-t-il de régions de décisions induites par une fonction discriminante linéaire? Donnez leurs équations.

4.3 (3 pts) Expliquez en français, précisément, mais dans vos propres mots la notion de frontière (ou surface) de décision. Donnez l'équation de la frontière de décision induite par une fonction discriminante linéaire. Comment nomme-t-on une telle forme géométrique? Quelle est la dimensionalité d'un tel objet?

4.4 (3 pts) Expliquez en français, précisément, mais dans vos propres mots la notion de séparabilité linéaire. A quoi applique-t-on cette notion (de quoi dit-on que c'est oui ou non linéairement séparable). Formalisez la notion par une écriture mathématique.

4.5 (4 pts) Tracez pour $d = 2$ et $d = 1$ un exemple de cas linéairement séparable et un exemple de cas non linéairement séparable (notez que cela donne 4 graphiques en tout). Pour chacun, tracez la meilleure frontière de décision linéaire possible. Indiquez clairement quel graphique correspond à quel cas, et indiquez par une légende la frontière de décision et les régions de décision.

4.6 (3 pts) Que se passe-t-il si on roule l'algorithme du perceptron original sur un cas linéairement séparable? Et sur un cas non-linéairement séparable?

5 Maximum de vraisemblance (25 pts)

5.1 (3 pts) A quoi sert le principe de maximum de vraisemblance (on s'en sert pour trouver quoi? pour quel genre de problème?).

5.2 (2 pts) Quelle relation y a-t-il entre le principe de maximum de vraisemblance et le principe de minimisation du risque empirique?

5.3 (3 pts) On considère un modèle de densité Gaussienne isotropique (sphérique). Quels sont ses paramètres et quelle est la dimension de chacun. Si on veut apprendre ces paramètres, combien de nombres réels va-t-on apprendre en tout?

5.4 (15 pts) Formulez la maximisation de la vraisemblance pour ce modèle et résolvez-la, en détaillant les étapes. (Vous pouvez répondre à cette question après la séparation horizontale ci-dessous et poursuivre dos de la page).

5.5 (2 pts) Si on avait choisi d'utiliser un modèle plus compliqué, pour lequel il n'y aurait pas de solution analytique simple, comment s'y prendrait-on pour trouver un maximum?

Réponse à la question 5.4 ici ...