FONDEMENTS DE L'APPRENTISSAGE MACHINE (IFT3395/6390)

Professeur: Pascal Vincent

Examen Intra

Mercredi 19 octobre 2011

Durée: 1h45

Seule documentation permise:
une feuille recto-verso format letter avec vos
propres notes/résumé de compréhension du cours.

Prénom:
Nom:
Code permanent:
IFT3395 ou IFT6390:
Programme d'études :

Veuillez répondre aux questions dans les zones de blanc laissées à cet effet.

Notations

Les notations suivantes sont définies pour tout l'examen, *là où elles ont un sens*:
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 (composantes): $x^{(i)} \in \mathbb{R}^d$: $x^{(i)} = (x_1^{(i)}, ..., x_d^{(i)})$. On suppose aussi qu'un algorithme d'apprentissage permet de trouver une fonction $f_\theta(x)$ dont les paramétrée par θ et qu'on peut appliquer à tout nouveau point de test x.

ATTENTION: on vous demande de respecter *scrupuleusement* la notation définie dans cet énoncé. Par exemple souvenez-vous qu'ici la cible est notée t (ou $t^{(i)}$ s'il s'agit du ième exemple).

1 Problèmes d'apprentissage et minimisation du risque (25 pts)

1.1 (5 pts) Dans le cadre général donné en introduction, expliquez brièvement dans vos propres mots la différence entre un problème de classification, un problème de régression, et un problème d'estimation de densité de probabilité.

1.2 (12 pts) Spécifiez pour *chacun* des types de problèmes: **a)** la nature de la cible t (quelle plage de valeurs elle peut prendre); **b)** la nature de la sortie que l'algorithme va calculer pour pour un point de test x; **c)** la fonction de perte (ou de coût) précise $L(z, f)$ typiquement utilisée pour ce type de problème:

1. Classification binaire

2. Classification multiclasse

3. Régression

4. Estimation de densité

1.3 (8 pts)

En respectant les notations définies pour l'examen:

a) donnez la formule générale du *risque empirique* (le coût moyen sur un ensemble de données)
$\hat{R}(\qquad\qquad) =$

b) Expliquez en Français dans vos propres mots en quoi consiste le *principe de minimisation du risque empirique*: que permet-il de trouver?

c) Donnez l'expression mathématique correspondant à la minimisation du risque empirique

d) Dans le cas de l'estimation de densité, la minimisation de ce risque porte aussi un autre nom, lequel?

2 Exercice de classification (25 pts)

On a affaire à un problème de classification à 4 classes. L'ensemble de données D_n contient $n = 1000$ points, dont 400 sont de la classe 1, 400 sont de la classe 2, 100 sont de la classe 3, et 100 sont de la classe 4. On suppose qu'on a créé 4 estimateurs de densité $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$, $\hat{f}_4$, et entraîné chacun uniquement sur les points d'une classe ($\hat{f}_1$ a été entraîné sur les points de la classe 1, $\hat{f}_2$ sur ceux de la classe 2, etc...).

a) Nommez 3 techniques différentes que vous connaissez avec lesquelles on aurait pu construire ces estimateurs de densité.

b) Pour un nouveau point de test x que l'on désire classifier, on obtient en appliquant ces 4 estimateurs de densité à ce point:

$$\begin{aligned} \hat{f}_1(x) &= 0.5 \\ \hat{f}_2(x) &= 1.0 \\ \hat{f}_3(x) &= 2.5 \\ \hat{f}_4(x) &= 1.5 \end{aligned}$$

Expliquez brièvement comment vous vous y prendriez pour calculer le vecteur des probabilités d'appartenance aux classes pour ce point x: $(P(t=1|x), P(t=2|x), P(t=3|x), P(t=4|x))$. **Calculez ce vecteur.**

Votre réponse:
$(P(t=1|x), P(t=2|x), P(t=3|x), P(t=4|x)) = (\quad , \quad , \quad , \quad)$

c) Quelle classe d'appartenance décidera-t-on pour ce point x?

d) Dans quel cas un tel classifieur sera-t-il qualifié de « classifieur de Bayes **naif** »?

3 K-plus proches voisins pour la classification (25 pts)

3.1 (4 pts) Expliquez brièvement mais clairement l'algorithme de **classification** des K plus proches voisins (K-ppv ou K-NN) pour un problème de classification à m classes. Expliquez précisément dans vos propres mots comment l'algorithme effectue une prédiction pour un point de test x.

3.2 (6 pts)

Écrivez sous la forme d'un pseudo-code la fonction de classification du 1 plus proche voisin (cas simple où $K = 1$) mais pour m classes. Vous pouvez faire appel à une fonction $d(a, b)$ qui retourne la distance entre deux points a et b.

ClassifPlusProcheVoisin(x,D_n): #doit retourner la classe prédite pour x.

3.3 (2 pts)

Quelle est la complexité algorithmique de K-ppv (nombre d'opérations élémentaire) pour le calcul de prédiction de classe pour un point de test x (en fonction de n, d, m, K).

3.4 (2 pts)

On considère pour l'instant D_n comme étant l'ensemble d'apprentissage (ou d'entraînement). Quel taux d'erreur de classification sur cet ensemble (l'erreur d'apprentissage ou d'entraînement) obtiendra-t-on pour K = 1? On rappelle que l'erreur d'apprentissage est obtenue en calculant l'erreur de classification sur chaque point de l'ensemble d'apprentissage comme s'il était 1 un point de test (mais sans pour autant l'enlever de l'ensemble, même pas temporairement). Pour cette question, on suppose que tous les $x^{(i)}$ de D_n sont différents les uns des autres.

3.5 (2 pts)

Quelle réponse (prédiction) le classifieur donnera-t-il pour tout point de test si $K = n$?

3.6 (5 pts)

Expliquez comment vous procéderiez pour choisir la valeur de K qui a les meilleures chances de donner une bonne réponse pour de futurs points de test (on suppose n assez grand mais pas infini). Donnez la procédure sous la forme d'un pseudo-code de haut niveau.

3.7 (4 pts)

Sur un graphique, tracez l'allure *typique* des courbes d'erreur d'apprentissage et de test (taux d'erreur de classification sur l'ensemble d'apprentissage (ou d'entrainement) et sur un ensemble de test ou de validation séparé) en fonction de K, qu'on s'attend à obtenir. Indiquez clairement par une légende quelle est la courbe d'apprentissage et la courbe d'erreur de test.

4 Classifieur linéaire (25 pts)

On considère l'utilisation d'un classifieur linéaire (ex: Perceptron) pour un problème de classification binaire (2 classes: +1 et -1).

4.1 (3 pts)

Donnez l'expression mathématique d'une fonction discriminante linéaire (il devrait y avoir x dans votre expression...):

$g(x) =$

ainsi que la fonction de classification correspondante (qui va donner une classe +1 ou -1):

$h(x) =$

4.2 (3 pts) Précisez quels sont les *paramètres* de la fonction g (apparaissant dans l'expression que vous avez donnée ci-dessus) qu'on va vouloir apprendre. Précisez leur nom ainsi que leur *dimensionalité*.

4.3 (3 pts)

Quelle est la complexité algorithmique (nombre d'opérations élémentaire) pour le calcul de prédiction de classe pour un point de test x (en fonction de n, d). Voyez-vous un avantage, ou un inconvénient par rapport au K-PPV plus proche voisin? Lequel et dans quel cas s'applique-t-il particulièrement?

4.4 (5 pts)

Donnez une définition en français puis une définition mathématique pour les ensembles suivants:

a) la région de décision de la classe positive:

b) la frontière de décision:

c) Comment appelle-t-on l'objet géométrique qui correspond aux frontières de décision obtenues avec un classifieur linéaire?

4.5 (3 pts) Quand dit-on qu'un ensemble de données D_n est linéairement séparable? Expliquez-le d'abord dans vos propres mots, puis donnez une expression mathématique pour caractériser cette propriété.

4.6 (3 pts) Illustrez sur un graphique un ensemble de données 2D ($x \in \mathbb{R}^2$) de classification binaire, **linéairement séparable**, comportant 3 points de la classe $+$ et 3 points de la class $-$ (les 3 points de chaque classe doivent être non alignés). Tracez en **trait plein** la frontière de décision qu'on obtiendrait avec l'algorithme du **1 plus proche voisin**. Hachurez ou grisez légèrement la région de décision des $+$. Tracez ensuite en **trait pointillé** un exemple de frontière de décision qu'on pourrait obtenir sur ce problème avec un algorithme de type **classifieur linéaire** (fonction discriminante linéaire).

4.7 (2 pts) Même question que la précédente, mais avec un ensemble de données **non linéairement séparable**.

4.8 (3 pts)

Nommez tous les algorithmes d'apprentissage que vous connaissez qui permettent d'apprendre une fonction discriminante linéaire.