FONDEMENTS DE L'APPRENTISSAGE MACHINE (IFT3395/6390)

Professeur: Pascal Vincent

Examen Intra

Mardi 17 février 2009

Durée: 1h45

Toute documentation papier est permise (livre, notes de cours, ...)

Prénom:

Nom:

Code permanent:

IFT 3395 ou 6390?

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): $x^{(i)} \in \mathbb{R}^d$: $x^{(i)} = (x_1^{(i)}, ..., x_d^{(i)})$

ATTENTION: on vous demande de respecter *scrupuleusement* la notation définie dans cet énoncé. Donc avant de recopier directement des formules de vos notes de cours, assurez-vous d'y substituer la bonne notation! Sans quoi on pourrait conclure que vous ne comprenez pas à quoi correspond ce que vous écrivez... donc prenez le temps de bien assimiler la notation ci-dessus. Par exemple souvenez-vous qu'ici la cible est notée t (ou $t^{(i)}$ s'il s'agit du ième exemple).

1 Exercice de classification (10 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...).
Pour un nouveau point de test x que l'on désire classifier, on obtient en appliquant ces 4 estimateurs de densité à ce point:

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

a) 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.

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

c) Comment s'appelle cette technique ou ce genre de classifieur?

2 Plus proches voisins (30 pts)

a) Expliquez brièvement mais clairement l'algorithme de classification des K plus proches voisins (K-ppv ou K-NN) pour un problème à m classes. Expliquez précisément dans vos propres mots comment l'algorithme prend sa décision pour un point de test x.

b) Quel taux d'erreur de classification sur l'ensemble d'entraînement (l'erreur d'entraînement) obtiendra-t-on pour K = 1? On rappelle que l'erreur d'entraînement est obtenue en calculant l'erreur de classification sur chaque point de l'ensemble d'entraînement comme s'il était 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.

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

d) Expliquez en détail 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).

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

f) 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.

g) Même question que la précédente, mais avec un sensemble de données **non linéairement séparable**.

h) Selon vous, que sont le principal avantage et le principal inconvénient d'un classifieur K-ppv par rapport à un classifieur linéaire (pour un grand ensemble d'apprentissage)?

3 Estimation de densité (30 pts)

Parmi les techniques d'estimation de densité, nous avons vu plus particulièrement:

- la technique d'estimateur de densité à fenêtres de Parzen, utilisant par exemple un noyau Gaussien.
- la technique consistant à apprendre les paramètres d'une distribution paramétrée spécifique, par exemple une Gaussienne.

Dans ce cas, les deux techniques utilisent la distribution Gaussienne (appelée aussi loi normale) multivariée (c.a.d. en dimension d), mais de manière assez différente.

a) Donnez l'équation d'une la densité de probabilité Gaussienne, dans le cas d'une Gaussienne multivariée *isotropique* (aussi parfois appelée sphérique).

b) Précisez en les nommant quels sont les paramètres libres d'une telle Gaussienne, ainsi que leurs dimensions.

c) Expliquez brièvement, dans vos propres mots, les principales différences que vous voyez entre les deux techniques mentionnées au début de la question.

d) Expliquez, pour chacune de ces deux techniques, ce qu'on peut considérer comme des *paramètres* du modèle (à apprendre en minimisant un risque empirique sur l'ensemble d'entraînement), s'il y en a, et ce qu'on considère habituellement comme des *hyper-paramètres*, s'il y en a (à choisir sur la base de la performance sur un ensemble de validation).

e) Expliquez brièvement, dans vos propres mots, en quoi consiste la phase "d'entraînement" pour chacune de ces deux techniques.

f) Que se passerait-t-il si on considérait le ou les *hyper-paramètres* que vous avez identifié ci-dessus, plutôt comme des *paramètres*, c.a.d. qu'on voudrait les apprendre sur l'ensemble d'entraînement? Vers quelle valeur tendraient-ils?

Quel effet cela aurait-il à votre avis sur l'erreur de généralisation?

g) Donnez pour chacune de ces deux techniques, l'expression de l'estimateur de densité obtenu après "entraînement", en utilisant la même notation que votre réponse à la question a) et la notation définie au début de l'examen:

Pour Parzen: $\hat{f}(x) =$

Pour l'approche paramétrique Gaussienne: $\hat{f}(x) =$

h) Faites un graphique illustrant un problème à une dimension ($x \in \mathbb{R}$) où les points tendent à être davantage concentrés en **3** endroits différents. Le but est d'illustrer la différence entre les estimateurs de densité obtenus avec chacune des deux techniques: tracez en trait plein l'estimateur de densité de Parzen obtenue avec un choix approprié d'hyper-paramètre, et tracez en trait pointillé l'estimation de densité obtenue avec l'approche paramétrique Gaussienne.

i) Faites le même genre d'illustration, mais pour un problème à 2 dimensions ($x \in \mathbb{R}^2$), où vous représenterez les densités, non pas par leur hauteur sur une 3ème dimension, mais par leurs "courbes de niveau". Vous pouvez vous contenter d'un niveau, indiquant la ou les régions où la densité estimée est relativement élevée. Là encore, utilisez des traits pleins pour Parzen, et des pointillés pour la méthode paramétrique Gaussienne.

4 Classification linéaire et Perceptron (30 pts)

On suppose qu'on a affaire à un problème d'apprentissage supervisé, où on considère une fonction de perte (ou fonction de coût) $L(f, (x, t))$ qui permet de calculer le coût associé à une entrée x pour laquelle on prédirait $f(x)$ alors que la vraie valeur cible est t.

a) Expliquez, dans vos propres mots, la différence entre risque empirique et risque espéré (ou erreur de généralisation).

b) Écrivez les expressions mathématiques générales du risque empirique et du risque espéré (en respectant les notations définies ci-dessus)

c) On suppose maintenant qu'on a affaire plus spécifiquement à un problème de classification binaire, avec des cibles $t \in \{-1, +1\}$. Exprimez la fonction de perte (ou de coût) associée aux erreurs de classification. Précisez, pour ce cas, l'expression mathématique pour le risque empirique, et pour le risque espéré.

d) On suppose en outre que f est une fonction paramétrée par un vecteur de paramètres θ. Expliquez dans vos propres mots le principe de minimisation du risque empirique (en quoi il consiste, qu'est-ce qu'il permet de trouver). Écrivez l'équation mathématique correspondante.

e) On précise maintenant la forme de ce classifieur: il va s'agir d'un classifieur linéaire (affine).
La fonction discriminante linéaire correspondante est donnée par

$$g(x) = w^t x + b$$

où on a utilisée la notation vectorielle (le ^{t} indique une transposée) de sorte que $w^t x$ indique un produit scalaire. Exprimez la fonction de classification binaire $f(x)$ correspondante.

$f(x) =$

f) Précisez quel est l'ensemble θ des paramètres de f

$\theta = \{ \qquad \}$

Pour chaque paramètre, indiquez de quoi il s'agit: comment l'appelle-t-on généralement en Français. Précisez s'il s'agit d'un scalaire, d'un vecteur ou d'une matrice ainsi que sa dimensionalité.

A combien de nombres réels cela correspond-t-il en tout?

g) Réécrivez le calcul de la fonction discriminante $g(x)$, sans utiliser la notation vectorielle, mais en détaillant le calcul avec les composantes des vecteurs.

$g(x) =$

h) En reprenant cette dernière expression détaillée de $g(x)$ et de $f(x)$, réexprimez de la manière la plus spécifique et détaillée possible l'équation de minimisation du risque empirique pour le coût de classification avec une fonction discriminante linéaire.

i) Expliquez brièvement dans vos propres mots la notion de paysage de coût (d'erreur) ou surface de de coût (d'erreur). Illustrez le concept par une figure, en indiquant clairement quelle quantité est représentée sur chaque axe. Expliquez en une phrase quelles sont les limitations de cette illustration en ce qui a trait à la dimensionalité des quantités impliquées.

j) Est-il possible de minimiser le risque empirique de h) par la technique de descente de gradient? Si oui, comment, si non, pourquoi?

k) L'algorithme du Perceptron permet d'apprendre les paramètres d'une fonction discriminante linéaire. Mais pour ce faire, il minimise un risque empirique légèrement différent. Donnez l'expression de ce risque légèrement différent (toujours en utilisant les notations de cet énoncé et en particulier la forme donnée ci-dessus pour $g(x)$)

l) Écrivez l'algorithme du Perceptron en ligne, toujours en respectant la notation de cet énoncé.

m) Par qui et en quelle année a été inventé l'algorithme du Perceptron?

n) Dans quel cas l'algorithme du perceptron est il garanti de converger vers une solution en un nombre fini d'itérations?