Fondements de l'Apprentissage Machine (IFT 3395/6390)
Examen Intra

Professeur : Pascal Vincent

Mercredi 14 novembre 2012
Durée : 2h00

Documentation permise : 1 feuilles recto/verso (format letter 8" 1/2 x 11") pour votre résumé de cours.

Prénom :
Nom :
Code permanent :
IFT3395 ou IFT6390 :
Programme d'études (et laboratoire d'attache s'il y a lieu) :

Le total de l'examen est sur 100pts. Veuillez répondre aux questions directement dans les zones de blanc laissées à cet effet. Répondez de manière concise, mais précise. **Bon examen !**

Notation

Pour toutes les questions, on suppose qu'on travaille avec un ensemble de données de départ comportant n exemples noté $D_n = \{z^{(1)}, \ldots, z^{(n)}\}$ avec, dans les cas supervisés, $z^{(k)} = (x^{(k)}, t^{(k)})$ où $x^{(k)} \in \mathbb{R}^d$ est l'entrée et $t^{(k)}$ est la cible correspondante. On vous demande de respecter scrupuleusement les notations de cet énoncé (c.a.d. ne vous contentez-pas de retranscrire des formules telles quelles mais adaptez les aux notations de l'énoncé si besoin).

1 Tâches et familles d'algorithme (18 pts)

Nommez les trois principales tâches classiques en apprentissage. Et pour chacune :
- (1 pt) précisez si il s'agit d'une tâche d'apprentissage supervisé ou non-supervisé.
- (1 pt) expliquez brièvement dans vos propres mots en quoi elle consiste.
- (1 pt) indiquez la nature de la cible.
- (3 pts) listez les algorithmes d'apprentissage que nous avons vu ou évoqués qui peuvent être utilisés pour la résoudre, en prenant soin de préciser pour chaque algorithme listé s'il est considéré comme une méthode "paramétrique" ou "non-paramétrique".

2 Mise en situation (22 pts)

Vous êtes embauché dans une entreprise qui fabrique 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'algorithmes de classification, 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.

A)

1. (2 pts) Êtes-vous d'accord avec lui ? Expliquez/justifiez votre réponse.

2. (4 pts) Si vous n'êtes pas d'accord, comment proposeriez-vous à votre collègue de procéder pour décider laquelle des variantes utiliser ? De plus 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 ? Expliquez le tout en détail dans vos propres mots.

B)

Après un examen plus approfondi, seul l'un des algorithmes candidats, appelons-le A, vous paraît suffisamment prometteur pour le problème considéré (les autres par exemple, vous paraissent souffrir d'une capacité trop faible). Cet algorithme possède un hyper-paramètre de contrôle de capacité $\lambda \in \mathbb{R}^+$. On notera $f_\theta \leftarrow A_\lambda(D)$ pour indiquer qu'on a appliqué l'algorithme A (avec une valeur d'hyper-paramètre λ), sur un ensemble de données d'entraînement D et qu'il a appris et retourné la fonction de prédiction f_θ. On suppose que $f_\theta(x)$ retourne 1 si elle juge que la personne sur l'image x est une personne connue/autorisée et 0 sinon. Les données de départ sont un ensemble D_n tel que défini au début de l'énoncé avec $n = 200000$.

1. (4 pts) Exprimez la fonction mathématique précise de calcul du taux d'erreur de classification encourue par une fonction f_θ sur un ensemble de donnée D.

   $\hat{R}(f_\theta, D) =$

2. (7 pts) Écrivez ci-dessous le pseudo-code détaillé d'une approche qui permettra de trouver, apprendre et retourner la fonction f_{θ^*} apprise en utilisant le meilleur hyper-paramètre possible parmi un ensemble fini Λ spécifié par l'utilisateur (c.a.d. pour $\lambda \in \Lambda$).

3. (5 pts) Dessinez les courbes d'apprentissage typiques que pourrait permettre de produire votre pseudo-code ci-dessus (si on l'étendait pour produire des graphiques). Prenez soin d'indiquer clairement ce qu'il y a sur chacun de vos axes, et d'indiquer à quoi correspond chacune de vos courbes (notamment en les nommant).

3 Sur-apprentissage, sous-apprentissage, capacité et sélection de modèle (12 pts)

On rappelle que la "capacité" d'un algorithme d'apprentissage correspond, informellement, à la taille, la "richesse" ou la "complexité" de l'ensemble de fonctions considérées parmi lesquelles il trouve sa fonction de prédiction.

Répondez VRAI ou FAUX à la gauche de chaque ligne (ou bien abstenez vous) : +1 pour une bonne réponse, -1 pour une mauvaise, 0 pour une abstention (le minimum pour l'exercice est 0/12, le maximum 12/12).

1. Le sur-apprentissage se traduit par un taux d'erreur très faible sur l'ensemble de validation.
2. Le sous-apprentissage se traduit par un taux d'erreur trop élevée, à la fois sur l'ensemble d'entraînement et sur l'ensemble de validation.
3. Lorsqu'on a le choix entre plusieurs algorithmes d'apprentissage, on devrait choisir celui qui parvient le mieux à apprendre les exemples sur lesquels il a été entraîné.
4. Plus nombreux sont les paramètres à apprendre, plus il y a de risque de sur-apprentissage.
5. Plus on a d'exemples pour l'entraînement, plus il y a de risque de sur-apprentissage.
6. Un classifieur de fenêtres de Parzen à noyau Gaussien avec une largeur de fenêtre σ trop élevée mène à du sur-apprentissage.
7. Plus un algorithme d'apprentissage a une capacité élevée, meilleure sera sa prédiction pour de nouveaux exemples de test.
8. Plus un algorithme d'apprentissage a une capacité élevée, moins il fera d'erreurs sur un ensemble d'entraînement compliqué.
9. L'algorithme du *Perceptron* a une capacité plus élevée que le *1-plus proche voisin.*
10. La capacité d'un algorithme d'apprentissage peut généralement se contrôler à travers la valeur de ses *hyper-paramètres.*
11. Un algorithme d'apprentissage avec une plus forte capacité (qu'un autre) a un plus grand biais et une plus petite variance.
12. Si on choisissait la valeur des *hyper-paramètres* qui produit l'erreur la plus petite sur l'ensemble d'entraînement (sur lequel on apprend les *paramètres*) cela mènerait toujours à choisir la valeur des hyper-paramètres donnant la capacité la plus élevée possible.

4 Fonction discriminante linéaire (24 pts)

Nommons v le vecteur de poids et c le paramètre de biais d'une fonction discriminante linéaire $g_{v,c}$ applicable sur des points $x \in \mathbb{R}^d$.

1. (2 pts) Combien de paramètres scalaires (nombres réels) la fonction $g_{v,c}$ possède-t-elle ?

2. (1 pt) Exprimez sous forme d'une équation simple (vectorielle ou matricielle) le calcul qu'effectue $g_{v,c}$ en fonction explicitement de ses paramères v et c et de son entrée x :

   $g_{v,c}(x) =$

3. (2 pts) Exprimez le calcul de $g_{v,c}(x)$ sous forme d'une équation détaillée (utilisant une somme) en fonction des composantes scalaires du vecteur x :

   $g_{v,c}(x) =$

4. (2 pts) Exprimez la fonction de décision ou classification correspondant à cette fonction discriminante :

   $f_{v,c}(x) =$

5. (2 pts) À quoi correspondent les *régions de décision* $\mathcal{R}_+$ et $\mathcal{R}_-$ de chacune des deux classes (+1 et -1) induites par la fonction discriminante $g_{v,c}$. Expliquez brièvement cette notion d'abord dans vos propre mots, puis donnez des expressions mathématiques définissant ces deux ensembles.

6. (2 pts) À quoi correspond la *frontière de décision* induite par $g_{v,c}$. Expliquez brièvement cette notion d'abord dans vos propre mots, puis donnez une expression mathématique définissant cet ensemble (en fonction des paramètres de $g_{v,c}$).

7. (1 pt) Comment nomme-t-on ce type précis de frontière de décision (induit par $g_{v,c}$) ?

8. (1 pt) Exprimez la fonction de perte de type "erreur de classification binaire"associée à la fonction $f_{v,c}$ pour un exemple (x,t) où $t \in \{-1,+1\}$ indique la vraie classe de x :

   $L(f_{v,c}, (x,t)) =$

9. (2 pts) Exprimez le calcul du risque empirique obtenu avec $f_{v,c}$ sur un ensemble de données D_n

   $\hat{R}(f_{v,c}, D_n) =$

10. (1 pt) Nommez un algorithme permettant d'apprendre les paramètres de $g_{v,c}$

11. (1 pt) De quoi dit-on qu'il est ou n'est pas linéairement séparable (à quoi s'applique cette notion) ?

12. (2 pts) Définissez mathématiquement la notion de séparabilité linéaire en utilisant la fonction discriminante $g_{v,c}$ et ses paramètres.

13. (1 pt) Quelle serait la valeur de $\hat{R}(f_{v,c}, D_n)$ si les paramètres v, c ont été appris avec l'algorithme que vous avez indiqué en 10. sur un ensemble D_n linéairement séparable ?

14. (2 pts) Dessinez un exemple 2D linéairement séparable, et la frontière de décision qui serait obtenue par un classifieur linéaire. Dessinez aussi la frontière de décision obtenue par un 1 plus-proche voisin.

15. (2 pts) Dessinez un exemple 2D non-linéairement séparable, ainsi que la frontière de décision qu'obtiendrait un 1 plus-proche-voisin.

5 Maximum de vraisemblance et classifieur de Bayes naïf (24 pts)

On considère un problème de classification binaire (2 classes) dont les entrées x ont d traits caractéristiques réels et les cibles t indiquent la classe par $+1$ ou -1. On va appliquer un classifieur de Bayes naïf. Soit $D_{train} = \{(x^{(1)}, t^{(1)}), \ldots, (x^{(n)}, t^{(n)})\}$ notre ensemble d'entraînement, on définit les deux sous-ensembles $D_+ = \{(x,t) \in D_{train} | t = +1\}$ et $D_- = \{(x,t) \in D_{train} | t = -1\}$. On notera n_+ et n_- le nombre d'exemples dans $D+$ et D_- respectivement.

1. (2 pts) Quelle hypothèse distingue un classifieur de Bayes naïf d'un classifieur de Bayes général ?

2. (3 pts) On choisit de modéliser chacune des composantes de x (conditionnellement à la classe) avec une unique Gaussienne. Écrivez la forme paramétrique que prendra chacune de ces densités Gaussiennes en nommant et précisant les dimensions de ses paramètres. On notera $\hat{p}_{c,i}$ la densité associée à la classe c (+1 ou -1) et à la composante (trait caractéristique) i, et on vous suggère d'utiliser ces mêmes indices c, i pour distinguer les paramètres de chacun de ces $\hat{p}_{c,i}$.

   $\hat{p}_{c,i} =$

3. (3 pts) Exprimez précisément et de manière détaillée en fonction des données du problème ci-haut, le problème d'optimisation (une maximisation) découlant du principe de maximum de vraisemblance qui permettra de trouver les paramètres optimaux de chaque $\hat{p}_{c,i}$.

4. (2 pts) De quel principe un peu plus général le principe du maximum de vraisemblance est-il un cas particulier ?

5. (4 pts) Résolvez ci-dessous ce problème d'optimisation et finalement donnez les expressions simples résultantes permettant de calculer ces paramètres optimaux.

6. (3 pts) En fonction des $\hat{p}_{c,i}$ dont on suppose que les paramètres auront été correctement "appris" (c.a.d. calculés en utilisant les expressions appropriées de la question précédente), et en fonction des autres données du problème, exprimez le **vecteur** de probabilité de classes $g(x)$ que le classifieur de Bayes calculera pour un nouvel exemple x.

$g(x) =$

7. (3 pts) On suppose désormais pour simplifier que toutes les variances des Gaussiennes valent 1. Donnez l'expression mathématique de la frontière de décision de ce classifieur de Bayes (en fonction notamment des paramètres appris).

8. (4 pts) Montrez que cette frontière de décision correspond à la frontière de décision d'une fonction discriminante linéaire.