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

Examen Intra

Automne 2015

Professeur : Pascal Vincent

Jeudi 15 octobre 2015
Durée : 2h00

— Seule documentation permise : 2 feuilles recto/verso (format letter 8" 1/2 x 11") pour votre résumé de cours.
— L'utilisation d'appareils électroniques n'est pas autorisée durant l'examen (à l'exception d'une montre pour connaître l'heure).
— 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 !**

Prénom :
Nom :
Code permanent :
IFT3395 ou IFT6390 :

Programme d'études :
Laboratoire d'attache (s'il y a lieu) :

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. Et dans le cas non-supervisé /. 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 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, $\widehat{f_2}$ 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 :

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

1. 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.

2. Quelle classe d'appartenance décidera-t-on pour ce point x ?

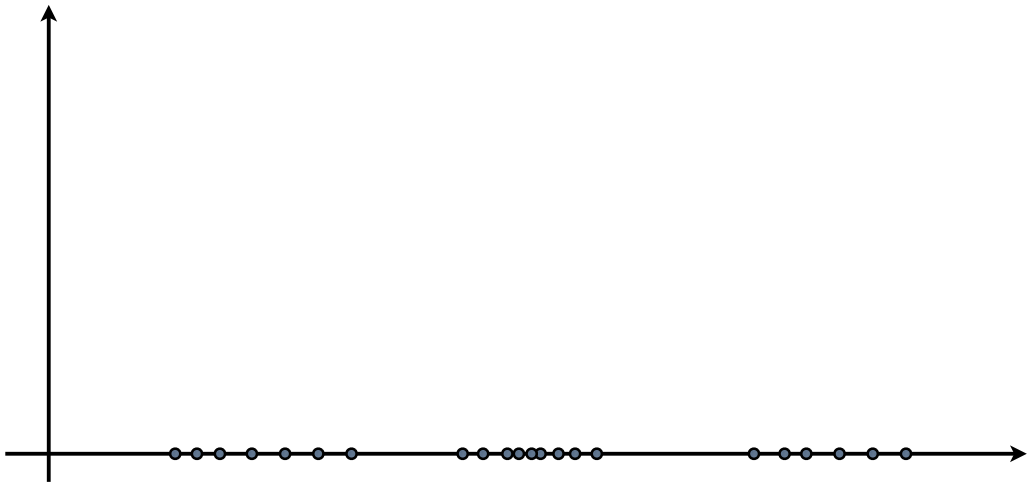
3. Comment s'appelle cette technique ou ce genre de classifieur ?

2 Estimation de densité (30 pts)

2.1 Allure de densité estimée obtenue avec divers algorithmes d'estimation

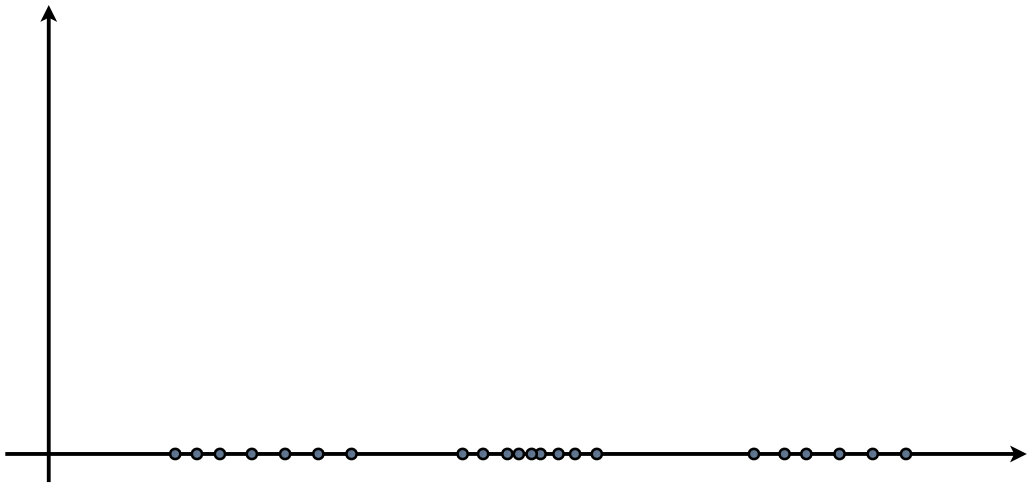
Sur le graphique ci-dessous qui représente un ensemble de données en dimension $d = 1$, tracez l'allure approximative de la densité estimée qu'on obtiendrait à l'aide des algorithmes d'estimation de densité suivants :

— en trait plein : un estimateur pramétrique Gaussien
— en trait pointillé : histogramme avec 8 subdivisions égales



Et sur le graphique suivant ci-dessous :

— en pointillé : Fenêtres de Parzen à noyau Gaussien avec un écart-type sigma trop grand
— en trait plein : Fenêtres de Parzen à noyau Gaussien avec un écart-type sigma vraiment trop petit
— en trait plein gras (appuyé) : Fenêtres de Parzen à noyau Gaussien avec un écart-type sigma approprié



2.2 Questions diverses

1. Pour chacun des algorithmes considérés ci-haut, nommez **tous** ses paramètres, et ses hyper-paramètres : (si un algorithme n'a pas de paramètres ou bien d'hyperparamètre, indiquez *aucun*).

   (a) Densité paramétrique Gaussienne :
   Hyper-paramètres :

   Paramètres :

   (b) Histogramme :
   Hyper-paramètres :

   Paramètres :

   (c) Fenêtre de Parzen à noyau Gaussien :
   Hyper-paramètres :

   Paramètres :

2. Comment devrait-on procéder pour choisir le meilleur estimateur de densité et les hyper-paramètres optimaux pour ce problème. Expliquez en détail dans vos propres mots.

3. Tracez ci-dessous l'allure des courbes d'apprentissage qu'on peut s'attendre à obtenir en variant l'hyper-paramètre des fenêtres de Parzen. Indiquez clairement les quantités sur vos axes. Identifiez clairement sur l'axe des abscisse, les zones qui correspondraient à du sur-apprentissge et les zones correspondant à du sous-apprentissage.

4. Écrivez la **formule détaillée** pour le calcul de ces courbes d'apprentissage (dans le cas des fenêtre de Parzen) c.a.d. comment on calcule ce qui est sur l'axe des ordonnées en fonction de ce qui est en abscisse.

3 Séparabilité linéaire (15 pts)

a) À quoi s'applique la notion de séparabilité linéaire : de quoi dit-on qu'il est ou non linéairement séparable ?

b) Parmi les définitions suivantes, lesquelles permettent de définir la notion "linéairement séparable" (entourez la ou les bonnes réponses) :

1. L'algorithme de classification linéaire considéré est capable d'atteindre 0 erreurs sur tout ensemble de donnée de classification binaire.
2. L'algorithme du perceptron stochastique (on-line) s'arrête au bout d'un nombre fini d'itérations.
3. On peut positionner un hyper-plan dans l'espace tel que tous les points d'une classe (parmi l'ensemble de donnée) sont d'un bord de l'hyper-plan et tous les points de l'autre classe sont de l'autre bord.
4. Tout classifieur linéaire fera 0 erreurs sur l'ensemble de donnée
5. Il existe une fonction discriminante linéaire dont le signe indique la classe de tout point généré par la distribution inconnue ayant généré les données
6. Il existe une fonction discriminante linéaire dont le signe indique la classe de tout point de l'ensemble de donnée
7. Soit $D_n = \{(x^{(1)}, t^{(1)}), \ldots, (x^{(n)}, t^{(n)})\}$ avec $x^{(i)} \in \mathbb{R}^d$ et $t^{(i)} \in \{-1, 1\}$
   $\exists w \in \mathbb{R}^d, b \in \mathbb{R}$ tels que $\forall (x, t) \in D_n$, $\text{sign}(w^T x + b) = t$
8. Soit $D_n = \{(x^{(1)}, t^{(1)}), \ldots, (x^{(n)}, t^{(n)})\}$ avec $x^{(i)} \in \mathbb{R}^d$ et $t^{(i)} \in \{-1, 1\}$
   $\exists w \in \mathbb{R}^d, b \in \mathbb{R}$ tels que $\forall (x, t) \in D_n$, $t(w^T x + b) < 0$
9. Soit $D_n = \{(x^{(1)}, t^{(1)}), \ldots, (x^{(n)}, t^{(n)})\}$ avec $x^{(i)} \in \mathbb{R}^d$ et $t^{(i)} \in \{-1, 1\}$
   $\exists w \in \mathbb{R}^d, b \in \mathbb{R}$ tels que $\left(\sum_{i=1}^{n} I_{\{(w^T x^{(i)} + b) t^{(i)} > 0\}}\right) = n$

c) Dessinez ci-dessous des exemples comportant une dizaine de points. Dans chaque cas, dessinez clairement les axes. Pour les cas linéairement séparables, dessinez également en plein la frontière de décision.

Dessinez ci-dessous un exemple 2D linéairement séparable	Dessinez ci-dessous un exemple 2D **non** linéairement séparable
Dessinez ci-dessous un exemple 1D linéairement séparable	Dessinez ci-dessous un exemple 1D **non** linéairement séparable (avec frontière de décision)

4 Régression avec prédicteur constant (15 pts)

On suppose qu'on a affaire à un problème de régression avec des cibles scalaires $t \in \mathbb{R}$. On va "apprendre" le prédicteur le pus simple possible : un prédicteur "constant" qui prédit toujours la même valeur c, quel que soit x :

$$f(x) = c$$

Son seul paramètre est donc $\theta = c$.

1. Quelle est la fonction de perte (coût) le plus souvent utilisée pour la régression ?

2. En partant du principe de minimisation du risque empirique, avec cette fonction de perte, exprimez (avec une équation **détaillée**) le problème d'optimisation spécifique qui permettra de trouver la valeur optimale c^* du paramètre.

   $\theta^* = c^* =$

3. Résolvez ce problème d'optimisation (écrivez toutes les étapes)

5 Classifieur de Bayes basé sur des densités Gaussiennes diagonales (30 pts)

La formule du calcul de densité pour une Gaussienne univariée (c.a.d. en 1 dimension) peut s'exprimer comme :

$$p(x) = \frac{1}{(2\pi)^{\frac{d}{2}}\sigma^d} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right)$$

La formule du calcul de densité pour une Gaussienne diagonale en dimension d peut s'exprimer comme :

$$\begin{aligned} p(x) &= (2\pi)^{-\frac{d}{2}} \left(\det(\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)^T \mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)^{-1}(x-\mu)\right) \\ &= (2\pi)^{-\frac{d}{2}} \left(\prod_{k=1}^{d} \sigma_k^2\right)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\sum_{k=1}^{d}\frac{(x_k-\mu_k)^2}{\sigma_k^2}\right) \end{aligned}$$

Où $\mathrm{diag}(\sigma_1^2, \ldots, \sigma_d^2)$ désigne la matrice diagonale qui a les valeurs $\sigma_1^2, \ldots, \sigma_d^2$ sur sa diagonale, et det est le déterminant.

1. Quels sont les paramètres d'une telle Gaussienne diagonale en dimension d tels qu'ils apparaissent dans cette formule (nommez et précisez-les) ? À combien de nombres scalaires cela correspond-t-il au total ?

2. Démontrez en détail que la densité d'une Gaussienne diagonale de dimension d peut s'exprimer comme le produit de densités Gaussiennes univariées.

3. Que peut-on en conclure concernant les d composantes d'une Gaussienne *diagonale* ?

4. On considère maintenant un classifieur de Bayes basé sur des Gaussiennes diagonales, pour un problème de classification à m classes. Indiquez précisément ce qui constituera son ensemble de paramètres (en nommant chacun et précisant ses dimensions) :

   À combien de nombres scalaires cela-correspond-t-il au total pour cet ensemble de paramètres ?

5. En fonction de ces paramètres, donnez l'équation détaillée de la **fonction de décision** (quelle classe est prédite) par ce classifieur :

6. Dans quel cas considère-t-on qu'on a un **classifieur de Bayes naïf** ? Soyez précis dans votre réponse.

7. Un classifieur de Bayes basé sur des Gaussiennes diagonales est-il un classifieur de Bayes naïf ? (Répondez OUI ou NON et justifiez votre réponse).

8. On suppose maintenant qu'on a un problème de classification à $m = 2$ classes ; et que les Gaussiennes modélisant chaque classe ont toutes deux des matrices de covariance diagonales

identiques (mêmes valeurs $\sigma_1^2, \ldots, \sigma_d^2$ sur leur diagonale). Démontrez en détail que la fonction de décision d'un tel classifieur de Bayes correspond à celle d'un classifieur (ou fonction discriminante) linéaire.

Indication : partez de la fonction de décision du classifieur de Bayes que vous avez indiqué plus haut, et montrez qu'elle correspond à la fonction de décision d'un classifieur linéaire.