le cnam

Sujet UE RCP208
Reconnaissance des formes et méthodes neuronales

Année universitaire 2018–2019

Examen 1ère session : 31 janvier 2019

Responsable : Michel CRUCIANU

Durée : 2h00

Seul document autorisé : 2 pages A4 recto-verso, manuscrites.

Les téléphones mobiles et autres équipements communicants (PC, tablette, etc.) doivent être éteints et rangés dans les sacs pendant toute la durée de l'épreuve. Calculatrices autorisées.

Sujet de 6 pages, celle-ci comprise.

Vérifiez que vous disposez bien de la totalité des pages du sujet en début d'épreuve et signalez tout problème de reprographie le cas échéant.

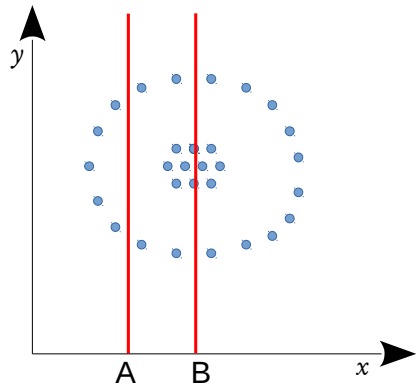
---

1. Expliquez brièvement pourquoi, lorsque le nombre de données (observations) N est très grand, le calcul de la densité par la méthode des noyaux présente un coût bien plus élevé que le calcul de la densité par un modèle de mélange gaussien avec un nombre faible de composantes. **(2 points)**

Correction:
Pour calculer la densité dans un point $\mathbf{x}$, par la méthode des noyaux il est nécessaire de calculer N noyaux (entre $\mathbf{x}$ et chacune des N données) alors que par un modèle de mélange seulement k lois normales, k étant le nombre de lois du mélange, supposé faible ($k \ll N$).

---

2. Considérons les données bidimensionnelles ($\in \mathbb{R}^2$) de la figure suivante :



   (a) Si une ACP centrée était réalisée sur ces données, les valeurs propres obtenues seraient proches l'une de l'autre ou très différentes ? Expliquez brièvement. (**2 points**)

   (b) Supposons que les données de la couronne externe forment une classe et les données centrales une autre classe. L'analyse factorielle discriminante est une méthode d'analyse adéquate dans ce cas ou non ? Justifiez brièvement. (**2 points**)

   (c) Indiquez un résultat probable de l'estimation d'un mélange de 5 lois normales bidimensionnelles sans contraintes sur les matrices de variances-covariance (`full` dans Scikit-learn) avec ces données, en dessinant des ellipses illustrant ces lois normales (une seule ellipse par loi). Et si on impose la contrainte d'avoir des matrices de variances-covariances diagonales (`diag` dans Scikit-learn) ? (**2 points**)

   (d) Quelle méthode de classification automatique serait appropriée pour ces données ? Justifiez brièvement. (**2 points**)

   (e) Deux observations incomplètes, A et B, pour lesquelles seule l'abscisse (axe x) est connue, sont représentée dans la figure par les traits horizontaux (l'ordonnée, ou l'axe y, étant à chaque fois inconnue). Quel serait le résultat d'une imputation par la moyenne dans ce cas ? Et celui d'une imputation par les 3 plus proches voisins ? Vous ne pouvez

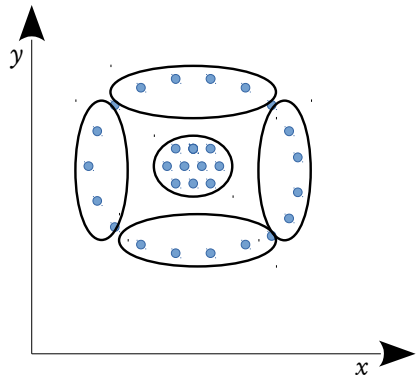
pas obtenir un résultat exact à partir de la seule illustration, expliquez seulement l'idée de la réponse. (**2 points**)

Correction:

(a) Le nuage d'observations étant relativement sphérique, les deux valeurs propres d'une ACP centrée seront proches l'une de l'autre. En effet, la dispersion des projections des observations sur une droite passant par le centre de gravité du nuage sera approximativement la même, quelle que soit l'orientation de cette droite.

(b) Le centre de gravité des données de la couronne externe est pratiquement le même que le centre de gravité des données centrales et les nuages sont relativement sphériques autour de ces centres de gravité, les projections des deux classes sur une droite passant par le centre de gravité du nuage seront superposées presque de la même façon, quelle que soit l'orientation de la droite. L'AFD a alors comme effet une perte d'information discriminante et est donc inadaptée dans ce cas.

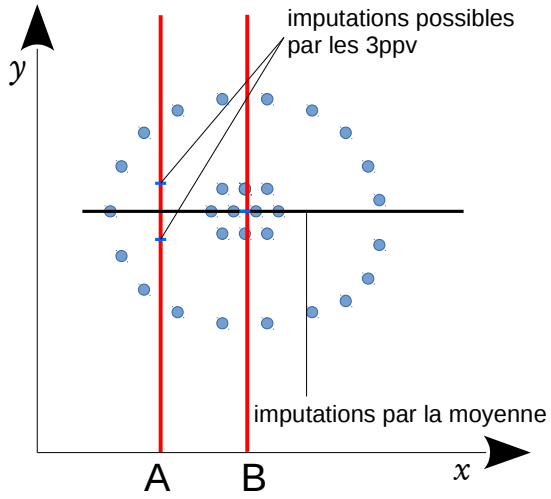
(c) Un résultat possible est présenté dans la figure suivante. La « spécialisation » des quatre gaussiennes se partageant les données de la couronne peut varier suivant l'initialisation de l'algorithme. Avec la contrainte `diag` (variances quelconques mais covariances nulles), le résultat sera proche de la figure ci-dessous.



(d) En raison de la présence de la couronne, le partitionnement par densité serait mieux approprié que le partitionnement en groupes compacts et bien séparés entre eux. Donc, parmi les méthodes vues en cours, la CAH avec le critère du lien minimum et la classification spectrale sont mieux adaptées que la CAH avec un autre critère d'agrégation ou que *K-means*.

(e) La figure suivante illustre les résultats de ces deux méthodes d'imputation. L'imputation par la moyenne donnera exactement le même résultat pour l'ordonnée de A et pour l'ordonnée de B. L'imputation par les 3 plus proches voisins (3ppv) sera pratiquement la même pour B. Pour A, en revanche, deux points dans la partie haute de la couronne

et 2 points dans la partie basse de la couronne seront pratiquement à la même distance sur x (seule coordonnée connue pour A) de A, donc suivant le résultat exact du calcul de proximité les 3ppv seront soit 2 points du haut et un du bas, soit deux points du bas et un du haut ; l'ordonnée obtenue pour A sera donc une des deux illustrées.



---

3. On veut résoudre un problème de classification à l'aide d'un perceptron (sans couche cachée).
   (a) Si le problème est linéairement séparable, le perceptron converge vers une solution (s'arrête lorsque tous les E^k sont nuls). Expliquez brièvement pourquoi cette solution peut ne pas être robuste. (**1 point**)
   (b) Si le problème n'est pas linéairement séparable, pourquoi le perceptron ne trouve pas une solution finale sur laquelle s'arrêter (il oscille entre plusieurs solutions possibles) ? (**1 point**)

Correction:

(a) Seuls les exemples en erreur contribuent à la correction des poids et l'algorithme s'arrêtera à la première solution trouvée (lorsque tous les E^k sont nuls). Donc rien ne garantit que cette solution sera robuste, c'est à dire qu'elle puisse donner de bonnes réponses pour des exemples non appris (c'est la notion de généralisation). (Avec la règle de Widrow-Hoff, tous les exemples participent à la mise à jour des poids et on obtient une solution qui est souvent plus robuste).

(b) L'algorithme du perceptron ne peut s'arrêter que lorsque tous les E k sont nuls, c'est à dire lorsqu'il classifie correctement l'échantillon d'apprentissage. Mais le perceptron

ne peut résoudre que des problèmes linéairement séparables. Donc si l'échantillon n'est pas linéairement séparable, le perceptron ne peut pas trouver une solution finale sur laquelle s'arrêter, il oscille alors entre plusieurs solutions possibles. (Avec la règle de Widrow-Hoff, on aboutit toujours à une solution acceptable)

---

4. On veut résoudre un problème de prédiction à l'aide d'un perceptron multicouches. Pour ce problème, on essaie de prédire une valeur en utilisant les 12 valeurs précédentes (comme pour le problème « sunspot », vu en TP). Le réseau de neurones utilisé possède une couche cachée de 10 neurones. La fonction d'activation des neurones cachés est une tangente hyperbolique.
   (a) Quel est le nombre de poids (biais ou seuil + poids) de ce réseau ? **(1 point)**
   (b) Quelle fonction d'activation utilisez-vous pour les neurones de la couche de sortie ? **(1 point)**
   (c) Pourquoi il est nécessaire de normaliser les entrées (que se passe-t-il si on a une entrée très grande) ? **(1 point)**

Correction:

(a) Quel est le nombre de poids (biais ou seuil + poids) de ce réseau ?
   — Nombre de poids de la couche d'entrée vers la couche cachée : 12 x 10
   — Nombre de poids de la couche cachée vers la couche de sortie : 10 x 1
   — Nombre de biais de la couche cachée : 10
   — Nombre de biais de la couche de sortie : 1

   Total : (12 x 10) + (10 x 1) + 10 + 1 = 141

(b) Quelle fonction d'activation utilisez-vous pour les neurones de la couche de sortie ?
   Pour un problème de régression il vaut mieux utiliser une fonction d'activation linéaire dans la couche de sortie.

(c) Pourquoi il est nécessaire de normaliser les entrées (que se passe-t-il si on a une entrée très grande) ?
   La fonction d'activation des neurones cachés est une tangente hyperbolique. Il est impératif que les entrées soient de moyenne pas trop loin de zéro et de variance pas trop loin de 1. Si on a une entrée très grande, elle fait saturer plusieurs neurones et bloque l'apprentissage pour cet exemple.

---

5. Expliquez, avec vos propres mots, pourquoi l'apprentissage des cartes topologiques fait décroître le paramètre T (la température) dans un intervalle. Expliquez comment se forme l'ordre topologique durant l'apprentissage. **(3 points)**

Correction:
Lorsqu'on utilisait une valeur de T fixée, on s'est aperçu qu'une valeur élevée de T favorisait la formation d'un ordre sur la carte mais que cette dernière ne parvenait pas à se déployer sur l'ensemble des données ; au contraire, une petite valeur de T permettait le déploiement de la carte sur les données, mais la carte obtenue n'était pas nécessairement bien ordonnée. Pour répondre à cet antagonisme, la procédure utilisée consiste à initialiser la température T à une valeur élevée, favorisant ainsi l'apparition de l'ordre, puis à la faire décroître progressivement au cours des itérations de minimisation, permettant à la carte de recouvrir peu à peu la distribution réelle des observations.
