le cnam

Sujet UE RCP208
Apprentissage statistique : modélisation descriptive et introduction aux réseaux de neurones

Année universitaire 2021–2022

Examen 1ère session : janvier 2023

Responsable : Michel CRUCIANU

Durée : 2h00

Seul document autorisé : 2 feuilles 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 4 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. Analyse en composantes principales (ACP) et analyse des correspondances binaires (AFCB).
   (a) Dans quel(s) cas il est indispensable d'employer une ACP normée (centrer **et** réduire les variables) et non simplement une ACP centrée (variables centrées) ? **(1 point)**
   (b) Expliquez en bref le rôle des aides à l'interprétation employées pour l'ACP. **(2 points)**
   (c) Quels sont les résultats d'une analyse des correspondances binaires si les deux variables sont parfaitement indépendantes entre elles ? **(1 point)**

   Correction :
   (a) Lorsque les variables ne sont pas directement comparables (nature différente, unités de mesure différentes) la réduction permet de les rendre comparables. Sans réduction les résultats dépendraient, par exemple, des choix arbitraires d'unités de mesure pour les différentes variables.
   (b) Il y a deux types d'aides qui caractérisent la qualité de représentation des observations (ou des variables initiales) sur les différents axes factoriels et respectivement l'impact des observations (ou des variables initiales) sur l'orientation des axes factoriels. Plus de détails dans le support de cours.
   (c) Si les variables étaient parfaitement indépendantes alors les profils de toutes les modalités devraient être **confondus** pour chaque variable. Dans la pratique, l'échantillon d'observation étant fini, les profils ne sont pas parfaitement identiques ; pour éviter d'interpréter des projections alors que les variables sont indépendantes, il est utile de réaliser un test d'indépendance des variables avant d'appliquer l'analyse des correspondances.

2. Donner deux avantages et un inconvénient de DBSCAN par rapport à k-means. **(2 points)**

   Correction : Avantages : nombre de groupes déterminé automatiquement, permet de trouver des groupes non-convexes, identification des points aberrants.
   Inconvénients : plus lent que k-means, méthode transductive (impossible de classer des points a posteriori), plus grand nombre d'hyperparamètres.

3. Avec n observations, décrites par seulement 2 variables quantitatives, nous avons le choix entre la méthode par noyaux et la méthode par modèle de mélange gaussien pour estimer une fonction de densité.
   (a) Si la densité estimée est uniforme dans un rectangle, laquelle des deux méthodes devrait donner de meilleurs résultats et pourquoi ? **(1 point)**
   (b) Laquelle des deux méthodes est déterministe, c'est à dire donne le même résultat à partir des mêmes n observations ? **(1 point)**
   (c) Laquelle des deux méthodes est la plus coûteuse pour déterminer la densité en un point précis ? Pour le modèle de mélange on considérera le modèle estimé au préalable. **(1 point)**

   Correction :

(a) La méthode par noyaux car il y a autant de noyaux identiques que d'observations (n), alors qu'avec le modèle de mélange il y a comparativement peu de composantes dans le mélange et les écarts entre les valeurs de densités estimées dans le centre d'une gaussienne et à sa périphérie sont élevés.

(b) La méthode par noyaux donne toujours les mêmes résultats à partir des mêmes observations, alors que le résultat de l'estimation du modèle de mélange dépend de l'initialisation des paramètres.

(c) Pour la méthode par noyaux il faut calculer n noyaux alors que pour le modèle de mélange seulement k ($\ll n$) noyaux (modèle déjà estimé), donc la méthode par noyaux est la plus coûteuse.

4. Quel est le critère optimisé par t-SNE et UMAP ? Comment est-il optimisé ? (**1 point**)

   Correction : t-SNE et UMAP minimisent la divergence de Kullback-Leibler entre la distribution des voisinages p_{ij} dans l'espace d'entrée et celle des voisinages q_{ij} dans l'espace d'arrivée. La minimisation se fait par descente de gradient.

5. Pourquoi n'utilise-t-on pas une similarité gaussienne dans l'espace d'arrivée (l'espace réduit) d'une projection avec t-SNE ou UMAP ? Justifier brièvement. (**1 point**)

   Correction : Utiliser une similarité gaussienne dans l'espace d'arrivée impose que les points "voisins" avec une similarité non-nulle soient très proches, car le gaussienne a une décroissance rapide et une queue courte. C'est le problème dit de « l'agglutinement ».

6. Dans quel(s) cas ignorer (c'est à dire supprimer) les observations à valeurs manquantes biaise les résultats de la modélisation, c'est à dire privilégie certaines combinaisons de valeurs des variables par rapport à d'autres ? (**1 point**)

   Correction : Dans tous les cas où le manque dépend de la valeur de certaines variables (observées ou non observées) : cas MAR et respectivement MNAR.

7. Sélection de variables.

   (a) Indiquez un avantage et un inconvénient de la sélection de variables par rapport à la réduction (linéaire) de dimension. (**1,5 points**)

   (b) Pourquoi la sélection par filtrage est moins coûteuse que la sélection *wrapper* ? (**1,5 points**)

   Correction :

   (a) Les variables sélectionnées font partie des variables initiales et gardent donc leurs significations, contrairement aux nouvelles variables obtenues par réduction (même linéaire) de dimension. En revanche, la réduction de dimension permet de chercher dans un espace de solutions plus grand que la sélection de variables, et de faire cette recherche plus efficacement.

(b) Dans une approche *wrapper*, pour évaluer chaque sélection il faut apprendre un modèle décisionnel sur les variables sélectionnées. Dans l'approche par filtrage, chaque sélection peut être évaluée directement (calcul direct d'un critère de qualité prédictive ou d'un critère simple de redondance entre variables sélectionnées).

8. Un réseau de neurones à une couche cachée à fonctions d'activation non linéaires et couche de sortie à fonction *softmax* est employé pour une tâche de classification à deux classes.
   (a) La frontière de séparation entre classes est-elle linéaire ou non linéaire dans l'espace des variables d'entrée ? Et dans l'espace des projections des observations sur la dernière couche cachée ? (**2 points**)
   (b) Y a-t-il des points communs entre la régularisation par « oubli » et la régularisation par « arrêt précoce » ? Expliquez. (**2 points**)
   (c) Pourquoi faut-il estimer l'erreur de généralisation par l'erreur sur des donnés de test, non utilisées pour l'apprentissage, plutôt que sur les données d'apprentissage ? (**1 point**)

Correction :

(a) La frontière de séparation entre classes est linéaire dans l'espace des projections des observations sur la dernière couche cachée. Avec des fonctions d'activation non linéaires dans la couche cachée, la frontière de séparation entre classes est en général non linéaire dans l'espace des variables d'entrée.

(b) Les deux méthodes de régularisation permettent de « décourager » les poids des connexions à prendre des valeurs éloignées de 0.

(c) L'estimation de l'erreur de généralisation sur les données d'apprentissage serait une estimation excessivement optimiste, car le modèle a été optimisé précisément sur l'échantillon d'apprentissage.