le cnam

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

Année universitaire 2021–2022

Examen 1ère session : janvier 2022

Responsable : Michel CRUCIANU

Durée : 2h00

Aucune communication n'est autorisée durant l'examen à l'exception des échanges avec les serveurs du Cnam et avec les enseignants de cette unité d'enseignement.

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. Après une analyse en composantes principales (ACP), la projection des observations sur les deux premiers axes principaux (supposés représenter ici 80% de la variance du nuage des observations) est celle indiquée dans la Fig. 1.
   (a) Y a-t-il des observations qui ont un impact excessif sur l'orientation du premier axe factoriel ? Si oui, lesquelles ? **(1,5 points)**
   (b) Pouvons-nous déterminer à partir de la projection de la Fig. 1 quelles observations sont bien représentées par les deux premiers axes ? **(1,5 points)**

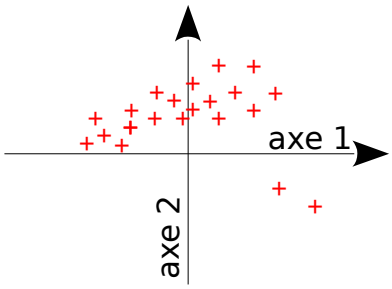

FIGURE 1 – ACP : projection sur les deux premiers axes

Correction :
   (a) Oui, les deux observations excentrées situées en bas à droite. **(1,5 points)**
   (b) Non. Bien que les deux premiers axes représentent 80% de la variance du nuage des observations, certaines observations peuvent être proches de l'origine dans cette projection mais éloignées de ce plan de projection, donc mal représentées par leurs projections sur ce plan. **(1,5 points)**

2. Quelle hypothèse présente dans DBSCAN concernant la répartition des données n'est pas présente dans ses successeurs (OPTICS et HDBSCAN) ? Justifier en quoi cette hypothèse peut affecter la performance du partitionnement. **(2 points)**

   Correction : DBSCAN suppose que la densité des groupes est identique pour tous les groupes. Si cette hypothèse n'est pas vérifiée, alors soit certains groupes peu denses seront séparés en points isolés (ε trop faible), soit certaines groupes plus denses seront agglomérés (ε trop élevé).

3. L'algorithme itératif EM est appliqué pour l'estimation d'un modèle de mélange de lois normales à partir de n observations.
   (a) Dans quelles situations les résultats obtenus à partir d'initialisations différentes sont très différents entre eux ? **(1,5 points)**
   (b) Pour quel nombre de composantes (lois normales) la vraisemblance est maximale ? Justifier en une phrase. **(1,5 points)**

Correction :

(a) Les résultats obtenus à partir d'initialisations différentes sont très différents entre aux lorsque les hypothèses sont erronées : distribution (par ex. uniforme) qui ne correspond pas du tout à un mélange de lois normales, ou nombre de composantes très mal choisi.

(b) Pour n composantes, chacune centrée sur une observation et « expliquant » donc parfaitement cette observation.

4. Quel est l'avantage de l'initialisation par ACP de l'algorithme t-SNE par rapport à l'initialisation aléatoire ? (**2 points**)

   Correction : Utiliser l'initialisation par ACP a deux intérêts : accélérer la convergence de l'algorithme de descente de gradient en proposant une initialisation des points qui capte déjà la structure globale ; et stabiliser les résultats en supprimant la stochasticité due à la répartition aléatoire des points initiaux.

5. Indiquez un avantage et un inconvénient de l'imputation par les k plus proches voisins ($k$ppv) par rapport à l'imputation par la moyenne. (**3 points**)

   Correction : Avantage : prise en compte de la distribution jointe des variables plutôt que de la distribution marginale de chaque variable.
   Inconvénients :
   - Coût potentiellement élevé de la recherche des $k$ppv dans le cas où il y a beaucoup d'observations.
   - Problème de pertinence dans le cas où il y a beaucoup de variables (malédiction de la dimension) : les $k$ppv ne sont pas nécessairement plus significatifs pour une observation que les autres observations.

6. Supposons que nous souhaitons sélectionner cinq variables parmi vingt variables explicatives quantitatives disponibles.
   (a) Entre une méthode incrémentale (ou par cooptation) et une méthode décrémentale (par élimination), laquelle est moins coûteuse ? Justifier brièvement. (**1,5 points**)
   (b) Expliquer brièvement ce qu'apporte la réduction de la redondance entre variables sélectionnées par rapport à la sélection sur le seul critère de qualité d'explication de la variable de sortie. (**1,5 points**)

   Correction :

   (a) Pour la méthode incrémentale comme pour celle décrémentale, le coût à chaque itération est proportionnel au nombre de variables restantes (à tester). Pour la méthode incrémentale le coût serait donc 20+19+18+17+16 alors que pour la méthode décrémentale 20+19+...+6. La méthode incrémentale est clairement moins coûteuse dans ce cas.

   (b) La réduction de la redondance permet de faire diminuer encore le nombre de variables sélectionnées (et donc la complexité du modèle) ou, à nombre constant de variables, de faire entrer dans le groupe de variables sélectionnées des variables *complémentaires* entre elles.

7. Un réseau de neurones à deux couches cachées à fonctions d'activation non linéaires est employé pour une tâche de classification multi-classes.
   (a) Quel codage doit être employé pour la variable expliquée dans le cas de la classification multi-classes ? Pourquoi ? (**1 point**)
   (b) Quel est l'intérêt de l'utilisation de la fonction d'activation *softmax* pour la classification ? (**1 point**)
   (c) Voyez-vous un intérêt à employer une (des) couche(s) cachée(s) linéaire(s) (fonction d'activation identité) dans un réseau de neurones ? Justifier brièvement. (**2 points**)

Correction :

(a) Un codage *one hot* (ou disjonctif) est préférable car il permet de représenter plus de deux classes sans introduire de relations particulières entre ces classes.

(b) La fonction *softmax* permet de normaliser les sorties d'un réseau et ainsi d'interpréter chaque sortie comme une probabilité d'appartenance du vecteur d'entrée du réseau à la classe correspondante. Elle est par ailleurs dérivable.

(c) Oui, **une** couche cachée linéaire avec un faible nombre de neurones permet de faire une réduction de la dimension des données en éliminant des directions de faible variance (comme l'ACP). **Plusieurs** couches linéaires qui se succèdent sont équivalentes à une seule. Bien entendu, si **toutes** les couches cachées sont linéaires le réseau est limité à des séparations linéaires.