le cnam

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

Année universitaire 2021–2022

Examen de rattrapage : avril 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. Analyse factorielle discriminante (AFD). **(5 points)**
   (a) Y a-t-il un danger à réduire la dimension des données par une ACP avant d'appliquer une AFD ? Justifier en bref. **(1 point)**
   (b) Dans quelles situations l'AFD et l'ACP appliquées aux mêmes données (sauf pour la variable « classe » à laquelle seule l'AFD a accès) trouvent les mêmes axes factoriels ? **(2 points)**
   (c) Y a-t-il un lien entre les axes discriminants trouvés par une AFD et les axes principaux trouvés par une ACP appliquée aux seuls centres de gravité des classes ? **(2 points)**

   Correction :
   (a) Oui, car les directions de plus faible variance, supprimées lors de la réduction de dimension par ACP, peuvent être discriminantes.
   (b) Lorsque les directions de plus forte variance sont aussi celles de plus forte séparation entre les classes, c'est à dire lorsque l'information de séparation entre les classes est l'information dominante présente dans les données.
   (c) Si la matrice des covariances empiriques totales $\mathbf{S}$ est la matrice identité, alors on constate que l'AFD maximise l'inertie inter-classes, c'est à dire réduit la dimension en conservant le maximum de variance entre les centres de gravité des classes, exactement ce que fait l'ACP des centres de gravité.

2. Classification automatique. **(3 points)**
   (a) On applique une classification *k-means* sur un jeu de données de n observations $\{x_1, x_2, \ldots, x_n\}$, avec $x_i \in \mathbb{R}^p$ un vecteur à p dimensions. Soit $y \in \mathbb{R}^p$ une nouvelle observation. Combien de distances $d(x, y)$ doit-on calculer pour savoir à quel groupe affecter y ? Justifier brièvement. **(2 points)**
   (b) Après application de DBSCAN sur un jeu de données, on observe que DBSCAN a retenu un seul groupe, contenant 95% des observations. Les 5% restants sont considérés comme des données aberrantes. Quelle hypothèse peut-on faire sur les valeurs des hyperamètres ε et `minVoisins` ? **(1 point)**

   Correction :
   (a) Chaque groupe est représenté par son centre m_j, avec $1 \leq j \leq k$. Le groupe auquel y est affecté est celui correspondant au centre le plus proche. Il est donc nécessaire de calculer les k distances $d(m_j, y)$.
   (b) Si toutes les données sont rassemblées dans un même groupe, alors la taille du voisinage ε est probablement trop élevée : tous les points sont voisins de tous les autres. Comme certaines données sont étiquetées comme aberrantes, la valeur de `minVoisins` est probablement suffisante. Si elle était très faible (1 ou 2), alors il y aurait peu d'observations isolées.

3. Estimation de densité. **(3 points)**

(a) La (log-)vraisemblance finale (atteinte à la fin de la convergence de l'algorithme EM) permet-elle de choisir le nombre m de composantes dans un modèle de mélange ? Justifier. **(1,5 points)**

(b) Comment procéder pour savoir si le nombre m de composantes d'un modèle de mélange est inadapté, sans appliquer un critère comme AIC ou BIC (qui exige d'explorer une plage large de valeurs) ? **(1,5 points)**

Correction :

(a) La (log-)vraisemblance finale augmente avec le nombre de composantes m pour atteindre sa valeur la plus élevée lorsque m est égal au nombre d'observations. Seule, elle ne permet donc pas de choisir le bon nombre de composantes. **(1,5 points)**

(b) Un moyen simple est d'appliquer 2-3 fois l'estimation du modèle de mélange avec des conditions initiales différentes : si la valeur de m est inadaptée aux données, les résultats varieront assez fortement. **(1,5 points)**

4. Réduction non-linéaire de dimension (**1 point**)

(a) Comment les points dans l'espace d'arrivée sont-ils initialisés lors d'une réduction de dimension par t-SNE ? Donner un inconvénient de cette initialisation.

Correction :

(a) Les points dans l'espace d'arrivée sont initialisés aléatoirement. Cela introduit une stochasticité élevée : les résultats de l'optimisation de t-SNE dépendent du hasard et de la distribution initiale des points y_i. Une alternative est d'initialiser les points d'arrivée par une analyse en composantes principales.

5. Sélection de variables. (**3 points**)

(a) Le critère de variance (sélection des variables de variance maximale) n'est pas lié à la qualité prédictive des variables choisies. Pourquoi il peut quand même avoir un effet favorable sur les performances du modèle prédictif qui emploie les variables sélectionnées ? (**1,5 points**)

(b) Pourquoi adopter des méthodes incrémentales ou décrémentales dans la sélection de variables ? (**1,5 points**)

Correction :

(a) Réduire le nombre de variables explicatives permet de réduire la complexité du modèle prédictif qui utilise ces variables et donc d'améliorer ses capacités de généralisation (à condition que les variables éliminées soient peu prédictives ou redondantes avec les variables conservées).

(b) Ces méthodes permettent de réduire l'espace de recherche qui serait, sinon, trop large : $O(C_m^k)$ pour sélectionner k variables parmi m.

6. Régularisation dans les réseaux de neurones multi-couches. **(5 points)**
   (a) La régularisation peut-elle être trop forte ? Justifier. (**1 point**)
   (b) Un réseau avec plusieurs couches cachées dotées de fonctions d'activation non linéaires est employé dans une tache de classification. Quelle sont les conséquences d'une prolongation de l'apprentissage par descente de gradient sur les poids des connexions, sur la frontière de discrimination et sur la capacité de généralisation du réseau ? (**2 points**)
   (c) On emploie un auto-encodeur avec plusieurs couches cachées dotées de fonctions d'activation non linéaires. Quelle est conséquence de l'utilisation d'une pondération α élevée pour le terme d'oubli ? (**2 points**)

Correction :

(a) Pour la plupart des méthodes, oui. Par ex., pour *weight decay*, un coefficient α trop élevé provoque une régularisation trop forte. En revanche, pour *early stopping*, arrêter l'apprentissage lorsque l'erreur de validation atteint son minimum ne peut pas provoquer une régularisation trop forte.

(b) Une prolongation de l'apprentissage peut produire une augmentation des valeurs absolues des poids de certaines connexions, avec comme conséquence la saturation de certains neurones (suivant les fonctions d'activation utilisées), l'augmentation du caractère non linéaire de la frontière de discrimination et une diminution des capacités de généralisation.

(c) Plus α augmente, plus les poids sont poussés vers 0, donc plus la fonction de transfert de l'auto-encodeur devient linéaire (ou affine) et plus son traitement s'approche d'une projection sur les composantes principales suivie d'une reconstruction linéaire des données.