le cnam

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

Année universitaire 2024–2025

Examen 2e session : avril 2025

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.

[4 points] Question 1 : Méthodes factorielles

1. A quoi servent les aides à l'interprétation dans une ACP ? **(2 points)**
2. Quel intérêt à l'inclusion de variables quantitatives dans l'analyse des correspondances multiples et comment procéder pour le faire ? **(2 points)**

Correction :

1. A savoir si une observation (ou une variable) est bien représentée par sa projection sur les axes (les contributions relatives) et si une observation (ou une variable) n'a pas un impact excessif sur l'orientation d'un axe (contributions absolues).
2. Cela permet de mettre en évidence des relations entre variables quantitatives et qualitatives qui caractérisent les observations (l'analyse factorielle des données mixtes permet également ce type d'analyse). Cela permet aussi d'aller au-delà des limites d'une analyse linéaire de ces variables quantitatives. Pour cela il est nécessaire de discrétiser les variables quantitatives en définissant des intervalles qui correspondront aux modalités.

[4 points] Question 2 : Classification automatique

1. Quel est le principe de k-means++ et pourquoi une telle initialisation est en général meilleure qu'une initialisation aléatoire uniforme ? **(2 points)**
2. Que calcule l'indice de Rand et pourquoi faut-il l'ajuster (pour obtenir l'indice de Rand ajusté) ? **(2 points)**

Correction :

1. k-means++ vise à mieux répartir des centres de groupes dans les données lors de l'initialisation, en les éloignant entre eux évitant d'avoir des groupes de données sans centre à proximité (si le nombre de centres n'est pas trop faible). Une telle initialisation permet une convergence plus rapide et vers une meilleure solution car les centres sont plus représentatifs des groupes de données dès l'initialisation.
2. L'indice de Rand permet de comparer des partitionnements sans avoir à faire appel à des numéros de groupes, en regardant dans quelle mesure les deux partitionnements regroupent les mêmes données et séparent les mêmes données. Il est en général utilisé sous une forme ajustée pour que la valeur obtenue puisse être interprétée quel que soit le nombre de groupes et d'observations.

[3 points] Question 3 : Estimation de densité

1. Quel est le principe des critères d'information pour le choix du nombre de paramètres pour les modèles de mélange ? **(2 points)**
2. Qu'est-ce qu'une loi multidimensionnelle « sphérique » et combien de paramètres a-t-elle ? **(1 point)**

Correction :

1. Les critères dits « d'information » visent à choisir le meilleur modèle en minimisant la somme entre la log-vraisemblance négative (qui diminue avec l'augmentation du nombre de paramètres) et une pénalité (qui augmente avec le nombre de paramètres).
2. Une loi multidimensionnelle est dite « sphérique » si sa matrice de variances-covariances est la matrice identité (multipliée par une valeur). Une telle loi dans l'espace de dimension d a $d + 1$ paramètres : d pour son espérance multidimensionnelle plus un pour sa variance sur chaque dimension (la valeur qui multiplie la matrice identité).

[3 points] Question 4 : Sélection de variables

1. Quel est l'intérêt des méthodes incrémentales ou décrémentales de sélection de variables ? Pourquoi sont-elles considérées sous-optimales ? **(2 points)**
2. Donnez un exemple (sous forme de dessin) de cas où la corrélation entre deux variables quantitatives est 0 mais l'information mutuelle entre les variables est (très) différente de 0. **(1 point)**

Correction :

1. Les méthodes incrémentales ou décrémentales permettent de réduire le coût de la sélection de variables en organisant une exploration itérative mais incomplète de l'espace de recherche. En effet, elles n'explorent pas entièrement l'espace de toutes les combinaisons possibles de variables, de meilleures solutions peuvent exister dans la partie non explorée de cet espace, c'est en cela qu'elles sont sous-optimales.
2. Voir les cas présentés sur la seconde ligne de la Fig.137 du support de cours.

[6 points] Question 5 : Réseaux de neurones

1. Quel est le lien entre l'encodage de la variable nominale « classe » à la sortie d'un réseau de neurones et le choix de la fonction d'activation softmax pour la couche de neurones de sortie ? Expliquer brièvement. (**2 points**)
2. Qu'apporte l'inertie (*momentum*) à la convergence de la descente de gradient ? (**1 point**)
3. Qu'est-ce qu'on entend par dilemme biais-variance ? Expliquer brièvement. (**2 points**)
4. Quel avantage présente un auto-encodeur non linéaire par rapport à un auto-encodeur linéaire ? (**1 point**)

Correction :

1. La fonction d'activation softmax normalise les activations de neurones de sortie (somme des activations égale à 1), ce qui est cohérent avec un encodage *one-hot* de la variable de sortie qui représente des classes mutuellement exclusives (classification « multi-classes »). [Si les modalités de la variable de sortie ne sont pas mutuellement exclusives (classification « multi-label », plusieurs étiquettes ou labels peuvent être valables pour chaque observations d'entrée) alors un encodage *one-hot* serait incorrect et l'emploi de softmax inapproprié.]

2. L'inertie permet d'accélérer la convergence sur un plateau de l'erreur (où le gradient de l'erreur est très faible) et de lisser la trajectoire (donc en général de l'accélérer) dans une vallée étroite (où le gradient de l'erreur changerait fortement de direction lors du passage d'un versant à l'autre).
3. Pour réduire le biais il faut élargir la famille de modèles et on est alors confrontés à une augmentation de la variance des performances par rapport à l'échantillon d'apprentissage (de nombre d'exemples fixé).
4. Lorsque le nuage d'observations est non linéaire, un auto-encodeur non linéaire à le potentiel d'ajuster mieux ce nuage qu'un encodeur linéaire, les codes obtenus perdent donc moins d'information et permettent de reconstruire (décoder) plus fidèlement les données encodées.