On utilise les cinq échelles formées lors du dernier devoir à partir des données `visatrans` pour faire de la segmentation ou profilage des clients. La base de données `visaechelles` contient les cinq variables `ech1`-`ech5` et un identifiant `id` ainsi que le sexe et la variable binaire représentant la possession de la carte VISA Première (`carvp`).

1. Créez une matrice de nuages de points des cinq échelles et commentez. Faites de même avec les composantes principales obtenues à partir de la matrice de corrélation des échelles. Combien de groupements distinguez-vous dans cette dernière?
2. Faites un regroupement hiérarchique des cinq échelles à l'aide de la méthode de Ward avec la dissemblance euclidienne (de base).
   (a) Produisez un graphique du critère R^2 semi-partiel en fonction du nombre de regroupements. Combien de groupes ce critère suggère-t-il?
   (b) Rapportez les statistiques descriptives par regroupement pour les variables `sexe`, `carvp` ainsi que les cinq échelles.
   (c) Interprétez les différents profils de clients ainsi obtenus.
   (d) Représentez graphiquement les groupes obtenus à l'aide d'une matrice de nuages de points sur les trois composantes principales des variables échelles.
   (e) Répétez cette analyse avec la méthode de liaison simple (plus proches voisins) et la méthode de liaison complète (voisins les plus éloignés). Est-ce que ces méthodes mènent à une meilleur segmentation? Ne considérez que l'option à trois groupes; justifiez adéquatement votre réponse.
3. En utilisation les moyennes des échelles pour les regroupements obtenus avec la méthode de Ward comme valeurs de départ pour l'algorithme des K moyennes, faites un regroupement avec trois groupes.
   (a) Représentez graphiquement les groupes obtenus à l'aide d'une matrice de nuages de points pour les trois composantes principales des échelles.
   (b) Est-ce que la méthode non-hiérarchique (K moyennes) améliore la segmentation? Utilisez le graphique pour argumenter quant à la qualité de la segmentation.
   (c) Expliquez pourquoi la segmentation n'est pas satisfaisante. Quel est le problème à l'origine de cette mauvaise performance?
4. Les composantes principales obtenues en faisant une décomposition en valeurs propres des échelles standardisées représentent les vecteurs propres de la matrice des corrélations et génèrent le même espace que les données standardisées. Refaites la segmentation en utilisant ces composantes principales pour la segmentation en lieu et place des échelles. Initialisez l'algorithme des K moyennes avec la solution à trois groupes de l'algorithme de Ward (mais avec la moyenne des segments pour les composantes principales).
   (a) Est-ce que la qualité de cette dernière est meilleure qu'avec les échelles?
   (b) Conclure quant à l'utilité de faire une normalisation/rotation des données a priori plutôt que d'utiliser les échelles (indication : en quoi les matrices de nuages de points des paires de variable diffèrent-elles?)

Compilé le 10/02/2021 à 15:37