Les données du devoir 1 ont été transformées en échelles suite à une analyse factorielle préliminaire. Le but de cet exercice est de faire de la segmentation ou profilage des clients. La base de données `visaechelles` contient les cinq échelles `ech1-ech5`, où les variables du Devoir 1 sont regroupées de la manière suivante :

```
ech1 = boppnl + kvunb + nbjdl + qcredl + opgnbl;
ech2 = endetl + gagel;
ech3 = itavcl + xlgmtl + ylvmtl;
ech4 = dmvtpl + kvunb;
ech5 = relat + age.
```

Grosso modo, `ech1` représente le niveau d'activité, `ech2` le niveau d'endettement, `ech3` la fortune, `ech4` le degré d'utilisation du compte (difficile à interpréter) et `ech5` l'ancienneté. La base de donnée inclut 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. Est-il nécessaire de standardiser les données pour la segmentation dans cet exemple? Justifiez votre réponse (en cas de doute, faites l'analyse avec ou sans standardisation et jugez de la qualité).
3. 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 à l'échelle originale des données.
   (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.
4. 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. Est-ce que la méthode non-hiérarchique (K moyennes) change sensiblement la segmentation? Utilisez un graphique pour argumenter quant à la qualité de la segmentation.

Compilé le 02/10/2021 à 16:08