Algèbre Linéaire et Analyse de Données

Examen Machine 2021 - 2022, Sujet 2, Durée : 1h20
Licence 2 MIASHS

Guillaume Metzler

Université de Lyon, Université Lumière Lyon 2
Laboratoire ERIC UR 3083, Lyon, France

guillaume.metzler@univ-lyon2.fr

Résumé

L'examen comporte deux exercices qui sont indépendants et qui nécessitent de reprendre les notions vus en cours depuis le début de l'année.
Le premier exercice porte sur la partie *Prise en Main de* R. Le troisième exercice porte sur l'*Analyse en Composantes Principales*.

Vous répondrez directement dans le fichier R qui accompagne ce sujet. Ce fichier R contiendra aussi bien le code que les commentaires qui permettra de répondre aux questions du sujet.

Quand vous aurez terminé, vous déposerez directement votre examen sur l'espace de dépôt prévu à cet effet, directement sur Moodle.

Exercice 1

On considère la matrice B définie par

$$B = \begin{pmatrix} 1 & 2 & 0 & 1 & 0 & 3 \\ 1 & 2 & 1 & 0 & 1 & 5 \\ 0 & 1 & 1 & 2 & 0 & 6 \\ 1 & 0 & 2 & 0 & 3 & 7 \end{pmatrix}$$

1. Déterminer les dimensions de B à l'aide de R.
2. Combien de valeurs singulières non nulles la matrice B possède-t-elle au maximum ?
3. Effectuer la décomposition en valeurs singulières de B.
4. Quelle est le rang de la matrice B ?
5. On rappelle que l'on peut écrire B sous la forme

   $$B = U\Sigma V^T.$$

   Rappeler le lien qui existe entre le premier vecteur singulier à gauche, la matrice B, la première valeur singulière et le premier vecteur singulier à droite. Expliciter ce lien directement sur R.
6. Définir des objets $C1$ et $C2$ qui contiendront respectivement les approximations de rang 1 et de rang 2 de la matrice B.
7. Déterminer la qualité de ces deux approximations.

Exercice 2

Ce dernier exercice a pour objectif de vous faire étudier un jeu de données à l'aide de l'Analyse en Composantes Principales.
Les données utilisées se trouvent directement dans le fichier R qui accompagne ce sujet. On présente rapidement les différentes variables, seules les huit dernières seront utilisées pour l'analyse :

- *titreCours* : le titre du cours.
- *idCours* : l'identifiant du cours.
- *inscription* : nombre de jours écoulés depuis votre inscription au cours.
- *progression* : votre progression sur le cours (en pourcentage).
- *moyenneDeClasse* : moyenne de la classe aux évaluations (en pourcentage).
- *duree* : durée estimée du cours (en heures).
- *difficulte* : difficulté estimée du cours (1 : facile... 3 : difficile).
- *nbChapitres* : nombre de chapitres.
- *nbEvaluations* : nombre d'évaluations dans le cours (comprend les quiz et les activités).
- *ratioQuizEvaluation* : proportion de quiz par rapport au nombre total d'évaluations (nombre d'évaluations : nombre de quiz + nombre d'activités).

Dans la suite, nous noterons X la matrice des données (ou encore la matrice de design). On rappelle que les individus sont représentés en ligne et que les variables sont représentées en colonne.

Préparation des données

1. Déterminer le barycentre du jeu de données.
2. Calculer l'écart-type de chaque variable.
3. Créer une matrice Z qui contiendra votre jeu de données *centré, réduit* et *normé*.
4. Déterminer la norme de Frobenius de la matrice Z.

Analyse du nuage des variables

Dans la suite de l'étude vous allez étudier le nuage des variables. Il sera judicieux de se reporter aux commandes graphiques qui ont été vues en cours pour faire la représentation et notamment pour l'interprétation.

1. Définir la matrice qui permet d'analyser le nuage des variables, on la notera K. Quelle est sa dimension ?
2. Diagonaliser la matrice K.
3. Quelle est lien entre la somme des valeurs propres de K et la norme de Frobenius de Z ?
4. Définir des objets $g1$ et $g2$ qui représentent les composantes principales des variables sur le premier plan factoriel (on désignera par $\mathbf{v}_1$ et $\mathbf{v}_2$ les vecteurs qui forment la base de ce premier plan factoriel).
5. Rappeler le lien qui existe entre *les composantes principales des variables* et *les axes principaux* lors de l'étude du *nuage des individus*
6. Représenter ces variables sur le premier plan factoriel, *i.e.* représenter le cercle des corrélations.
7. Quelle est la quantité d'information préservée lors de la projection des données sur le premier plan factoriel ?
8. Au regard du graphique obtenu, quel sens pourriez-vous donner aux axes $\mathbf{v}_1$ et $\mathbf{v}_2$? Justifiez votre réponse en indiquant les variables initiales qui ont permis votre interprétation.
9. Mise à part graphiquement, comment auriez-vous pu également faire cette interprétation ?
10. Quelles sont les variables les mieux représentées et les variables les moins bien représentées ?