Model-based learning

Projet sous le logiciel R - 2018-2019

L'objectif du projet est de créer un package R permettant de réaliser un algorithme de co-clustering pour des données quantitatives.

L'algorithme Le modèle considéré sera le modèle des blocs latents gaussien, estimé à l'aide d'un algorithme EM variationnel (VEM).
Votre fonction de co-clustering devra prendre en entrée :

- le jeu de données,
- le nombre de clusters en ligne et en colonne,
- le nombre d'initialisation de l'algorithme VEM (en spécifiant une valeur par défaut),

En sortie, l'algorithme devra retourner :

- les probabilités a posteriori pour chaque individu d'appartenir à chaque cluster en ligne, idem pour les colonnes,
- les partitions estimées par maximum a posteriori,
- les proportions, moyennes et variance de chaque co-clusters,
- la valeur du critère ICL.

Une seconde fonction `plot` associée au résultat de votre première fonction devra permettre de représenter la matrice de données co-clusterisée.

Le package Vos fonctions R devronta être incorporée à un package R.
Ce package comportera donc deux fonctions : une pour réaliser le co-clustering et une pour la représentation graphique. Comme tout package R il devra comporter une aide pour ces fonctions, incluant un exemple d'utilisation (un exemple simple qui soit rapide d'exécution).
Pour la création du package R, vous pourrez suivre ce document pour savoir comment procéder :
http://www.agrocampus-ouest.fr/math/livreR/faire_pkg_R.pdf

Applications

- Tout d'abord vous testerez votre modèle sur simulation et le comparerez au package `blockclusters`.
- Vous testerez ensuite la capacité du critère ICL à choisir le bon nombre de co-clusters.
- Vous appliquerez ensuite votre algorithme de co-clustering sur un jeu de données réelles de votre choix.

Livrable

- un package R (.tar.gz).
- un court rapport présentant le modèle, l'utilisation de votre package ainsi que les applications demandées ci-dessus.