TP+Projet: Co-clustering

Le projet est à réaliser par binome ou seul, A rendre le 23 janvier à 12:00 au plus tard. Attention, une partie du sujet de l'examen s'appuiera sur ce projet.

---

1 TP

Dans ce projet on s'appuiera d'abord sur le travail intitulé "Scénario" qui permet de mettre en pratique quelques méthodes de visualisation telles que ACP, AFC, AFCM, MDS et de classification telles que kmeans, CAH, NMF.

Les données, l'objectif et ces méthodes sont décrites dans le document disponible dans `http://wikistat.fr/pdf/st-scenar-explo-spam.pdf`. Il s'agit de 58 variables qui sont observées sur 4601 messages dont 1813 pourriels (spams). Les données sont disponibles dans le répertoire `http://wikistat.fr/data`. La variable binaire **Spam** est présente à titre illustratif mais qui sera utile dans la visualisation et la comparaison des méthodes.

1. Exécuter toutes les étapes décrites dans le document.
2. Proposer d'autres méthodes répondant aux mêmes objectifs.
3. Faire une synthèse de tous ces résultats.
4. Proposer une autre normalisation et mesurer son impact sur les résultats.
5. Sachant que la matrice des données est sparse et que le co-clustering serait très approprié.
   (a) Utiliser le package `Blockcluster`[1] avec le modèle approprié. Justifier votre réponse.
   (b) Visualiser les classes de messages.
   (c) Compléter votre analyse en termes de visualisation et de clustering en utilisant d'autres méthodes disponibles dans le package python `scikit-learn`[2].
   (d) Même question en utilisant le package `Coclust`[3]. Commenter vos résultats.
   (e) Investiguer le package `biclust` et particulièrement la méthode `BCQuest` sur la matrice codée en catégories. Commenter vos résultats.
   (f) Faire une analyse synthétique de vos résultats.

2 Projet

Dans le site `https://www.kaggle.com/datasets` plusieurs tables données sont disponibles.

1. Créer un compte et choisir deux tables parmi celles proposées.
2. Décrire ces deux tables et préciser l'objectif que vous vous êtes fixé en terme de (co)-clustering et visualisation.
3. Faire une analyse exploratoire des données.
4. Visualiser ces données avec plusieurs méthodes appropriées.
5. Choisir 3 méthodes de clustering, comparer les résultats obtenus.
6. Choisir 3 méthodes de co-clustering appropriées.
7. Faire une conclusion.

---

[1] `https://https://cran.r-project.org/web/packages/blockcluster/vignettes/blockcluster_tutorial.pdf`
[2] `http://scikit-learn.org/stable/`
[3] `https://pypi.python.org/pypi/coclust`