# Probabilité, Statistique et théorie de l'information

## Projet : Partionnement de données

Aurélien Decelle

April 23, 2020

**Abstract**

L'objectif du projet est de réaliser un algorithme de partitionnement de données basé sur l'inférence Bayésienne et de mesurer la quantité d'information à l'aide d'objects fondamentaux de la théorie de l'information. L'idée est de se confronter à un article de recherche, de savoir le lire et le comprendre, être capable d'écrire l'algorithme, et d'avoir ensuite un regard critique sur ce travail. Le projet sera donc séparé en deux parties. La première consistera à être capable de comprendre l'article et de reproduire l'algorithme. La seconde partie vous permettra de vous attaquer à d'eventuelles autres données et d'autres algorithmes de façon libre : des suggestions vagues seront proposées mais ce sera à vous de décider dans quelle direction regardée.

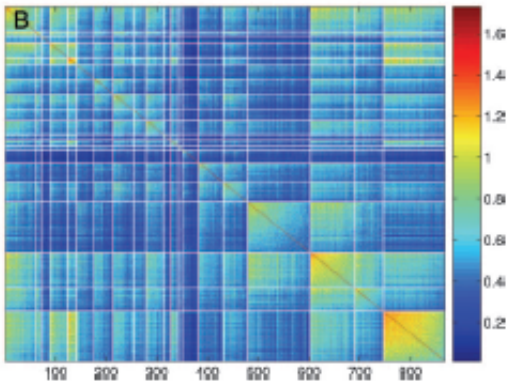


Figure 1: Matrice de similarité (information mutuelle) extraite de l'article

## 1 Introduction

La partitionnement de données est une tâche d'apprentissage omni-présente de nos jours effectué par le biais de divers algorithmes. Le principe du partitionnement est de regrouper un certain nombre de données selon des groupes permettant de mettre en valeur

des propriétés communes de ces données. Le partitionnement est souvent fait à l'aide d'une "distance" entre les différents objets à regrouper. En particulier, ces algorithmes sont particulièrement intéressants lorsqu'ils permettent de révéler une structure cachée au sein de ces données. Les exemples d'applications concrètes sont nombreuses, cela va de l'identification de communautés au sein par exemple de réseaux sociaux jusqu'à l'identification de fonctions particulières dans les réseaux métaboliques.

Ce projet s'intéressera au partionnement des données de façon non-supervisée (on ne connaît pas à priori si les données appartiennent intrinsèquement à un groupe ou non) et permettra de réaliser un algorithme de partitionnement basé sur les corrélations entre les données.

## 2 Matériel et description

Le projet s'articule autour de la méthode décrite dans l'article suivant[1]. Ce travail décrit un algorithme élaboré à partir de notions de théorie de l'information. Le projet est composé de deux parties, la première partie (la plus importante) devra répondre à un ensemble de question (cf ci-dessous) ainsi qu'à la réalisation de l'algorithme décrit dans l'article. Des jeux de données sont mis à disposition sur le site internet du cours[2] afin de pouvoir tester votre code. La seconde partie consistera au développement de certains points particuliers en relation avec l'algorithme.

## 3 Première partie

### 3.1 Questions de compréhension

Le modèle décrit dans l'article commence par définir $p(C|i)$,

1. à quoi correspond cette quantité ? De même pour la quantité $p(i|C)$.
2. En regardant l'équation (3), que représente la quantité $I(C, i)$ ?
3. Que se passerait-t-il si on ne fait que maximiser $\langle s \rangle$ ?
4. Expliquer alors pourquoi on cherche à maximiser $\langle s \rangle$ tout en minimisant $I(C, i)$.

### 3.2 Dérivation de l'équation (7) —

On va chercher à retrouver l'équation (7) à partir de l'équation (5).
**Premièrement**, on va s'intéresser au terme $I(C, i)$. On se souviendra tout d'abord que $P(C) = \sum_i p(C|i)p(i)$. Dans un premier temps, calculer la dérivée de $p(C)$ par rapport

[1] http://www.pnas.org/content/102/51/18297.full.pdf?with-ds=yes
[2] https://www.lri.fr/ adecelle/

à $p(C'|k)$. Montrer que l'on obtient pour la dérivée de $I$ :

$$\frac{\partial I}{\partial p(C'|k)} = \frac{1}{N}\left(\log(\frac{p(C'|k)}{p(C')}) + 1 - \sum_i \frac{p(C'|i)p(k)}{\sum_j p(C|j)p(j)}\right) \tag{1}$$

**Deuxièmement (questions bonus)**, nous allons regarder le terme $\langle s\rangle$. En notant que

$$p(i|C) = \frac{p(C|i)p(i)}{\sum_i p(C|i)p(i)} \tag{2}$$

Calculer la dérivée de $p(i|C)$ par rapport à $p(C'|k)$. Montrer que l'on obtient alors pour la dérivée de $\langle s\rangle$ :

$$\frac{\partial\langle s\rangle}{\partial p(C'|k)} = p(k)(1-r)s(C') + p(k)\sum_{s=1}^{r} s(C'; i^{(s)} = k) \tag{3}$$

**Finalement (pour tous)** Montrer que l'on obtient en prenant $p(i) = 1/N$, l'équation suivante :

$$p(C|i) = p(C)\exp\left[\frac{1}{T}\left(\sum_{r'} s(C; i^{(r')}) - (r-1)s(C)\right)\right] \tag{4}$$

où il manque la constante de normalisation. La constante de normalisation peut-être trouvée à l'aide de calculs supplémentaires, cela dit nous nous contenterons de considérer à partir de maintenant l'équation (7) comme référence.

### 3.3 Similarité pour $r = 2$ —

Pour la suite nous utiliserons une mesure de similarité dépendant de seulement deux variables $s(i, j)$. Rappeler la forme que prend l'équation (1) et (7) dans le cas $r = 2$.

### 3.4 Jeux de données

Afin de tester votre algorithme, les jeux de données suivants ont été générés artificiellement. Ces jeux de données correspondent à des points dans l'espace Euclidien générés par une mixture de Gaussiennes. Pour ces données, il vous est fourni la matrice de similarité correspondant ici à l'opposé de la distance euclidienne : $-||\vec{x} - \vec{y}||^2$.

- $N = 300$ données correspondant à des points dans l'espace euclidien 2D. La matrice de similarité des données se trouve sur le site du cours. Dans ce jeu de données (que vous pouvez également visualiser, par exemple à l'aide de gnuplot[3][4]), trois clusters d'environ $\sim 100$ points chacun sont présents et étiquetés. Vous pouvez ainsi vérifier que l'algorithme fonctionne correctement.

[3] http://www.gnuplot.info/

[4] par exemple par la commande : gnuplot 'file.dat' matrix with image

- $N = 300$ données, seule la matrice de similarité est donnée ici. Saurez-vous réussir à trouver combien de clusters sont présents dans ce jeu de données ?

**Finalement**, vous pouvez appliquer l'algorithme sur les deux jeux de données suivant :

- Variation du cours en bourse de 501 entreprises de l'indice américain SP500. Le jeu de données vous est fourni sous la forme d'une matrice correspondant à la mesure de similarité (cf l'article qui précise comment sont construites les matrices). Vous pouvez vérifier une partie de la cohérence de vos résultats à l'aide des informations se trouvant sur le site du projet. Sur le site du cours vous trouverez en plus des fichiers permettant l'identification des entreprises et une classification par secteur.
  - sp500_names.d : noms des 501 entreprises
  - sp500_matType.d : matrice reliant chaque entreprise à certains secteurs/sous-secteurs
  - sp500_TypeNames.d : noms des secteurs/sous-secteurs etc.
  - sp500_data.d : données "bruts" : cours de chaque entreprise

  Voir l'article pour plus de détails.
- Notation de 500 films (note comprise entre 1 et 6) par un total de plus de 70000 utilisateurs.
  - movie_name.d : noms des 500 films
  - movie_labels.d : matrice reliant chaque film à un ou des genres
  - movie_typename.d : noms des genres
  - movie_data.d : données "brutes" : matrice film-utilisateur contenant les votes de chaque utilisateur pour les films. Une entrée à zéro signifie que l'utilisateur n'a pas donné de vote pour un film.

Une façon de représenter le degré de clustering visuellement est d'afficher la matrice de similarité (à l'aide de gnuplot par exemple, cf plus haut) avec les lignes et les colonnes soigneusement permutées pour mettre les éléments d'un même cluster les uns à côté des autres. Pour permuter les lignes et colonnes d'une matrice, il faut utiliser une matrice de permutation P. Cette matrice est remplie de zéro sauf aux indices correspondants aux indices des lignes/colonnes à bouger. Par exemple, si vous voulez effectuer la permutation suivante :

$$\sigma = \begin{bmatrix} 1 & \to & 2 \\ 2 & \to & 4 \\ 3 & \to & 3 \\ 4 & \to & 1 \end{bmatrix} \tag{5}$$

il faut écrire la matrice de permutation suivante :

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \tag{6}$$

et ensuite effectuer l'opération suivante : $P^{-1}MP$ où $M$ est la matrice dont vous voulez permuter les indices. On note que $P^{-1}$ est donnée par la transposée de $P$.

Pour chacune de ces analyses, les moyens pour illustrer les résultats sont laissés au libre arbitre de chacun.