Méthodes Bayésiennes et processus de Dirichlet appliqués au Clustering

Mostafa Bouziane

Encadrant : Clément Elvira
Equipe Sigma à CRIStAL

11 mai 2017

Sommaire

C'est quoi le Clustering ?

Objectif : Structuration des données

- On ne dispose que d'exemples et le nombre de classes et leurs natures n'ont pas été prédéterminés.

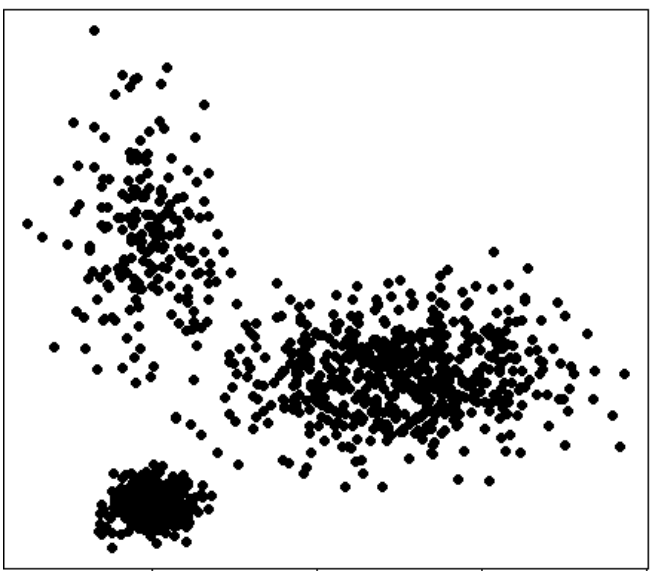


FIGURE 1 – Avant le Clustering

C'est quoi le Clustering ?

Objectif : Structuration des données

- L'algorithme doit découvrir par lui même la structure plus ou moins cachée des données.

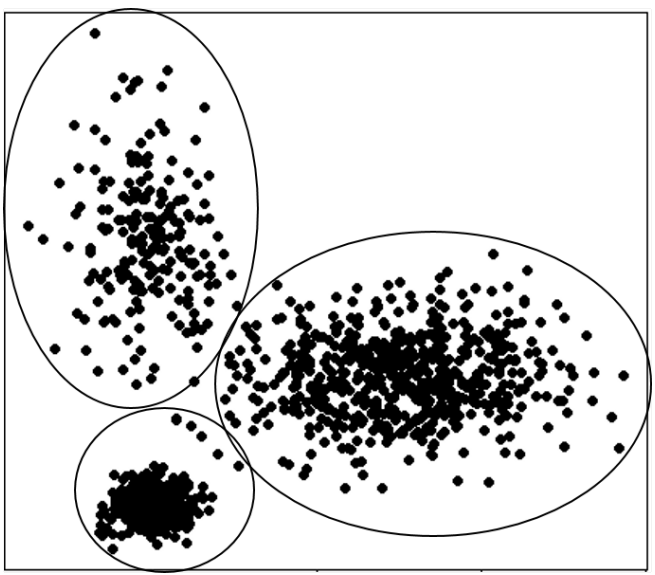


FIGURE 2 – Après le Clustering

Problématique

K nombre de clusters fixé

Algorithmes pour lesquels on fixe K au début :

- K-means
- K-Medoides
- Clustering Hiérarchique

Problématique

K nombre de clusters fixé

Algorithmes pour lesquels on fixe K au début :

- K-means
- K-Medoides
- Clustering Hiérarchique

Choix du nombre K

Comment choisir le nombre K de classes pour notre système de données ?

Méthodes Bayésiennes

Définition

C'est des méthodes qui permettent de déduire la probabilité d'un événement en se basant sur des événements déjà évalués.

Méthodes Bayésiennes

Définition

C'est des méthodes qui permettent de déduire la probabilité d'un événement en se basant sur des événements déjà évalués.

- Formule de Bayes : $p(A/B) = \frac{p(B/A)p(B)}{p(A)}$

Méthodes Bayésiennes

Définition

C'est des méthodes qui permettent de déduire la probabilité d'un événement en se basant sur des événements déjà évalués.

- Formule de Bayes : $p(A/B) = \frac{p(B/A)p(B)}{p(A)}$

Echantillonneur de Gibbs

C'est une méthode MCMC (Monte-Carlo à Chaînes de Markov), utilisée avec les méthodes bayésiennes pour l'échantillonage.

Méthodes Bayésiennes

Echantillonneur de Gibbs

C'est une méthode MCMC (Monte-Carlo à Chaînes de Markov), utilisée avec les méthodes bayésiennes pour l'échantillonage.

- La spécifité de cette méthode par rapport aux autres MCMC, c'est qu'elle découpe une distribution en plusieurs probabilités conditionnelles.

Clustering de Dirichlet

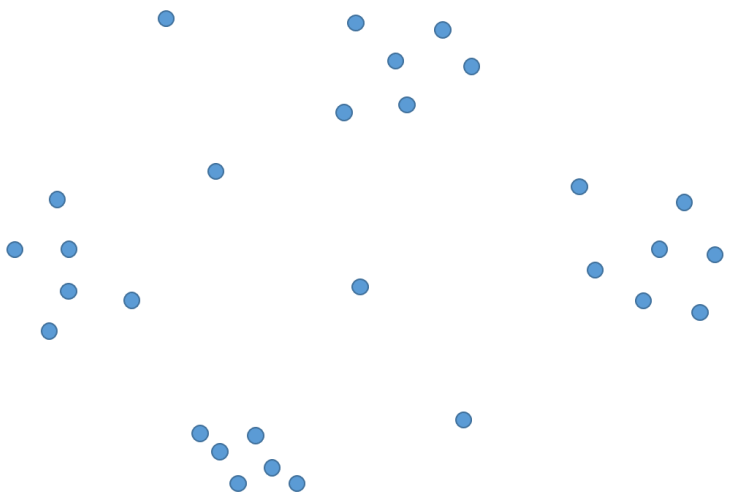


FIGURE 3 – Données à l'état initial assignées tous au même cluster

Clustering de Dirichlet

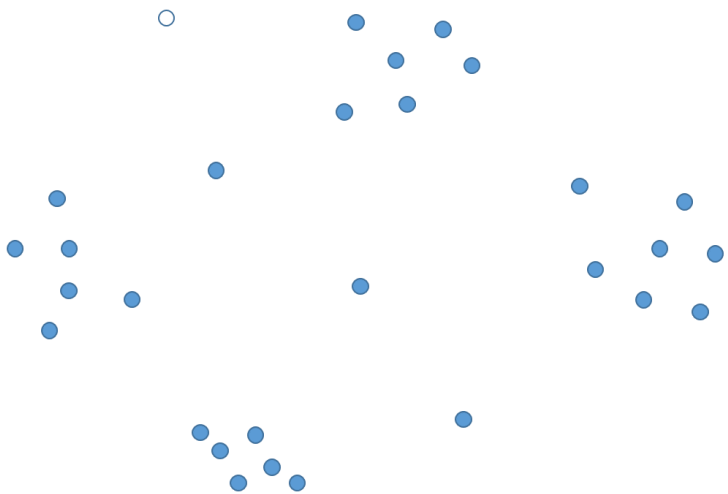


FIGURE 4 – On envlève la première observation du cluster

Processus de Dirichlet

Probabilité conditionnelle

On va se baser sur cette formule pour le Clustering :

$$\theta_{n+1}/\theta_1, \theta_2, ..., \theta_n \sim \frac{\alpha}{\alpha+n} H + \frac{1}{\alpha+n} \sum_{i=1}^{m} n_k \delta_{\theta_i^*}$$

Clustering de Dirichlet

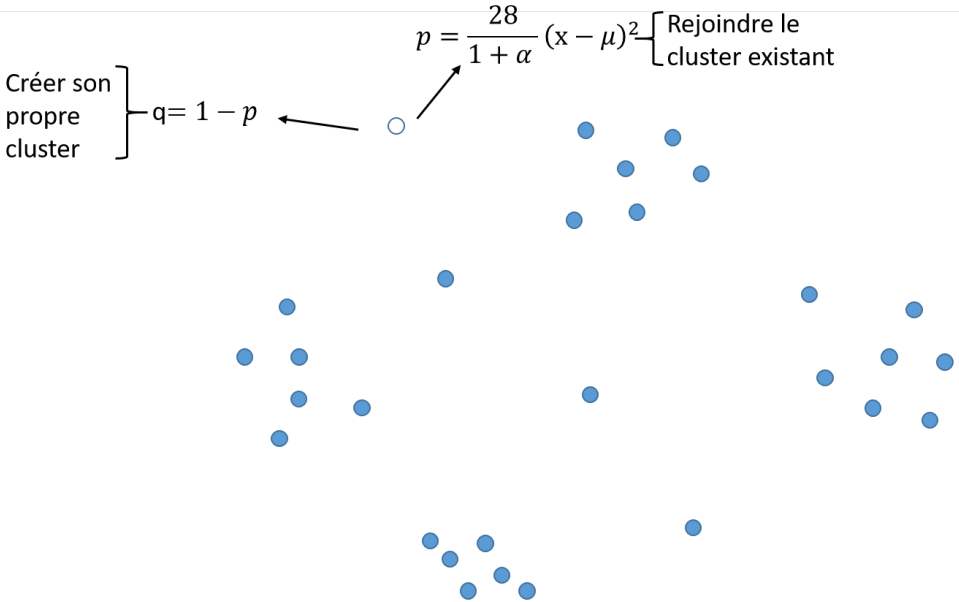


FIGURE 5 – On calcule les deux probabilités p et q

Clustering de Dirichlet

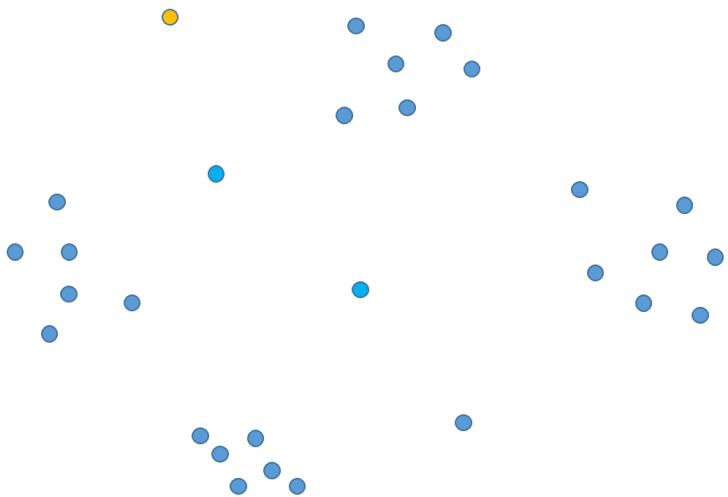


FIGURE 6 – Le point est assigné à un nouveau cluster

Clustering de Dirichlet

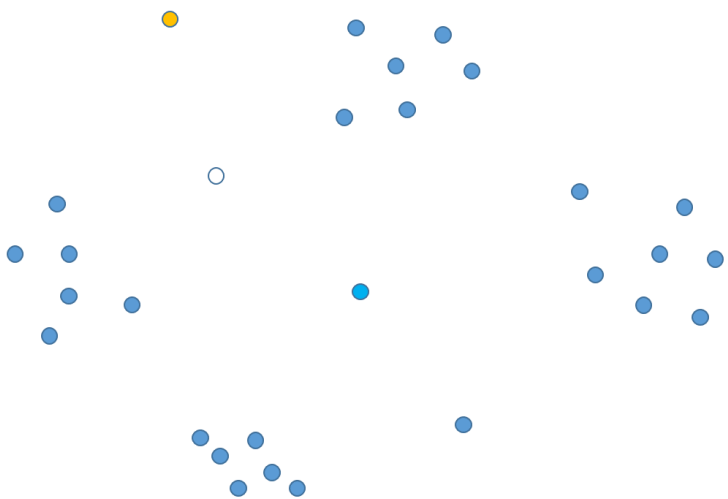


FIGURE 7 – On enlève l'observation de son cluster

Clustering de Dirichlet

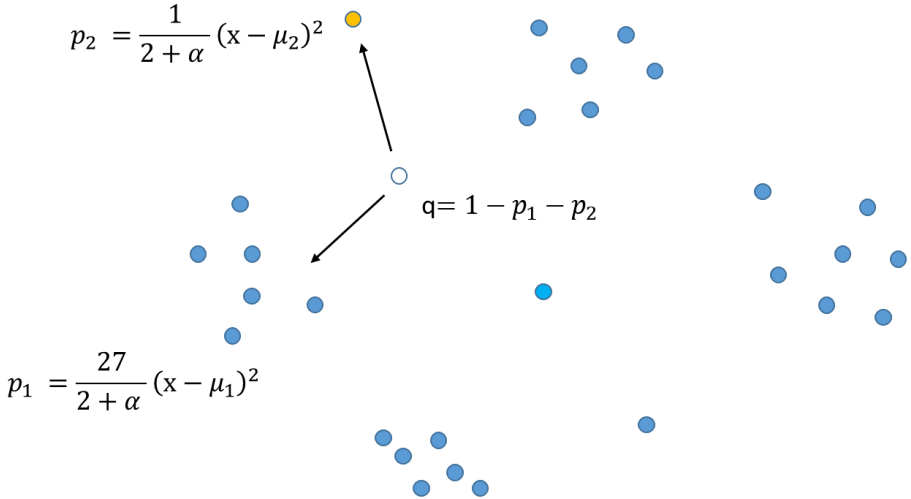


FIGURE 8 – On calcule les trois probabilités p_1, p_2 et q

Clustering de Dirichlet

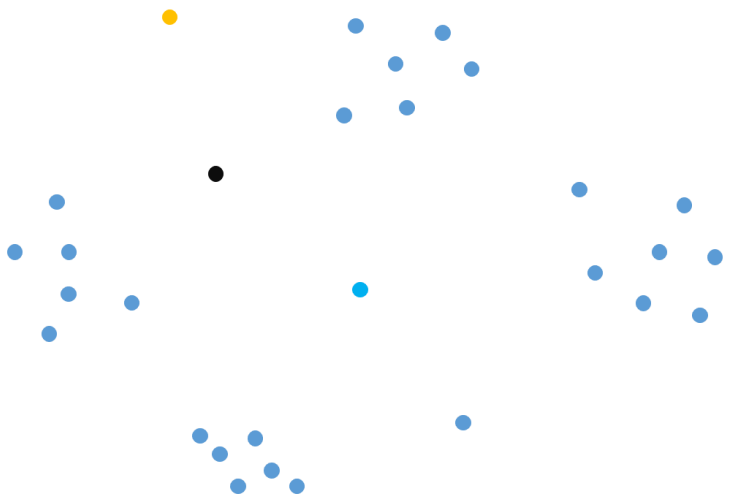


FIGURE 9 – Le point est assigné à un nouveau cluster

Clustering de Dirichlet

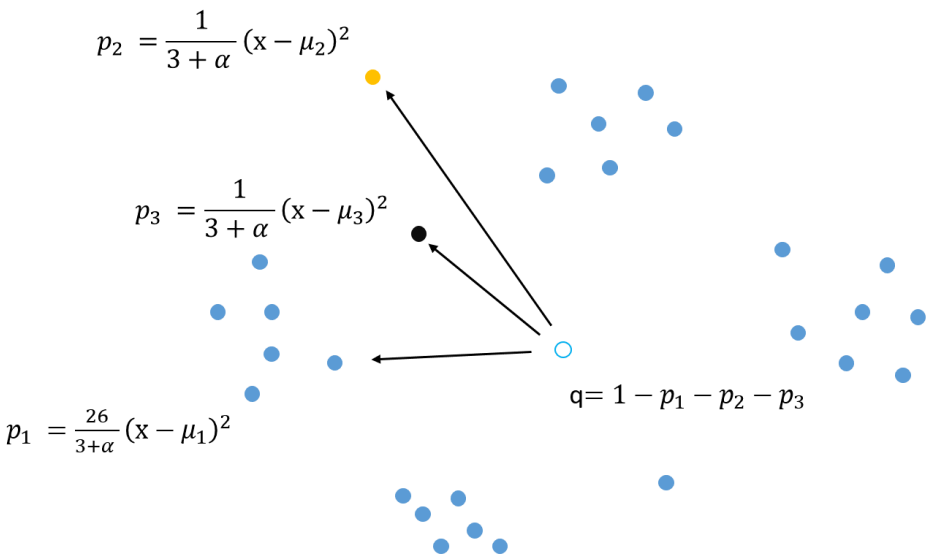


FIGURE 10 – Calcul des probabilités

Clustering de Dirichlet

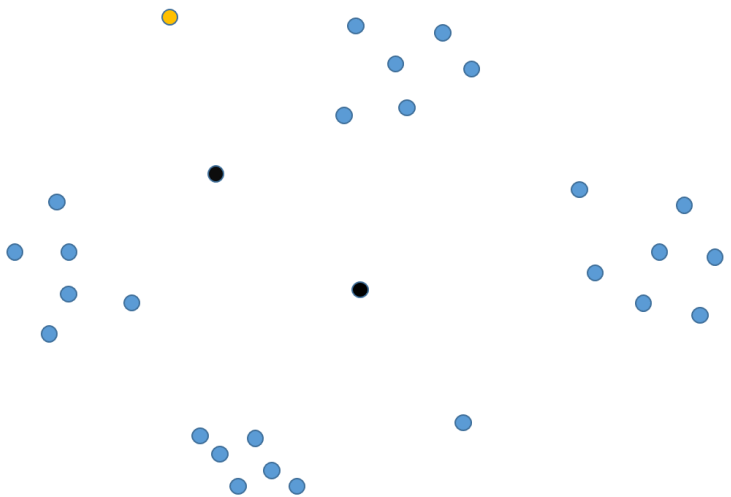


FIGURE 11 – Le point est assigné au cluster noir

Clustering de Dirichlet

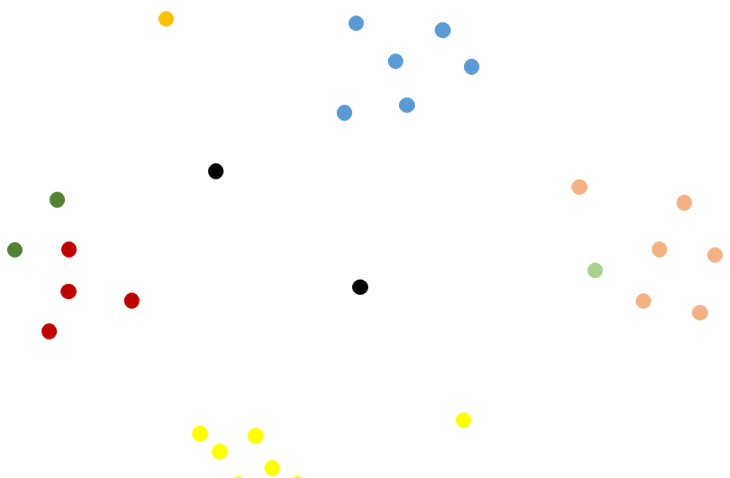


FIGURE 12 – On refait le même processus en tout point

MAP : Maximum à posteriori

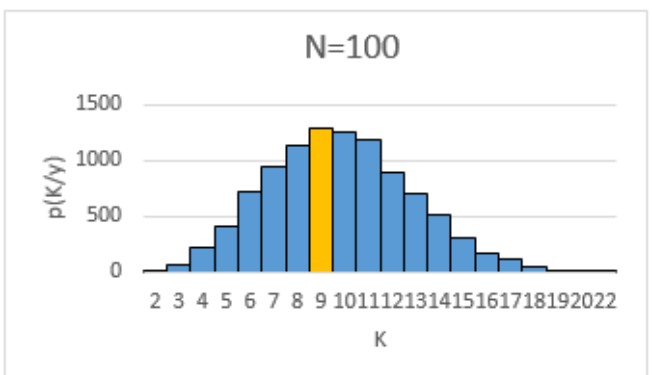


FIGURE 13 – Estimation pour 100 observations et 10 000 itérations de l'échantillonneur de Gibbs

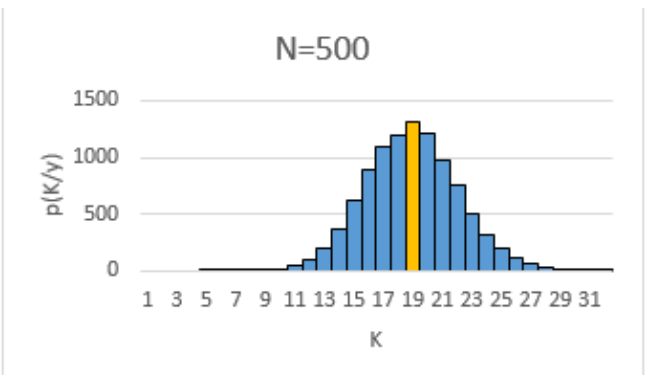


FIGURE 14 – Estimation pour 500 observations et 10 000 itérations de l'échantillonneur de Gibbs

MAP : Maximum à posteriori

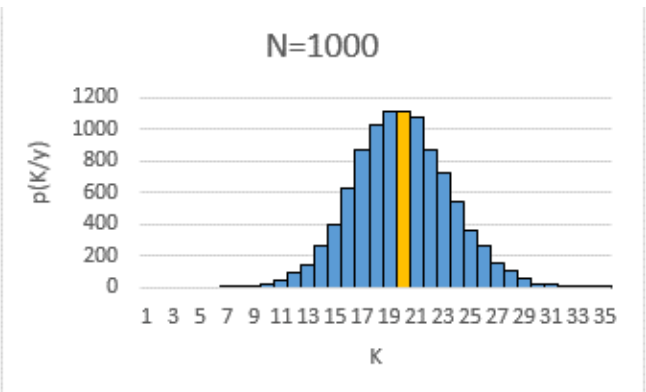


FIGURE 15 – Estimation pour 1000 observations et 10 000 itérations de l'échantillonneur de Gibbs

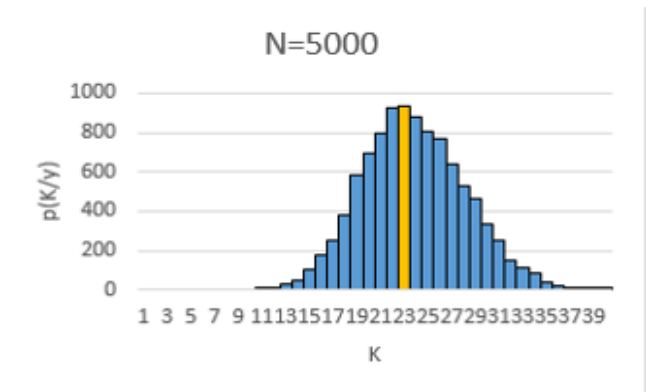


FIGURE 16 – Estimation pour 5000 observations et 10 000 itérations de l'échantillonneur de Gibbs

Applications

Marketing

Trouver des groupes de consommateurs similaires (mêmes produits consommés, mêmes comportements) dans une grande base de données.

1

1. www.linkedin.com/pulse/geo-clustering-new-dawn-market-segmentation-somshekhar-das

Applications

Biologie

Classification des animaux, plantes...

2

2. www.pinterest.com/nikkitten94/plants/

Applications

Assurance, domaine bancaire

Identifier des logements, des marchés, des cours de boursiers, les fraudeurs...

3

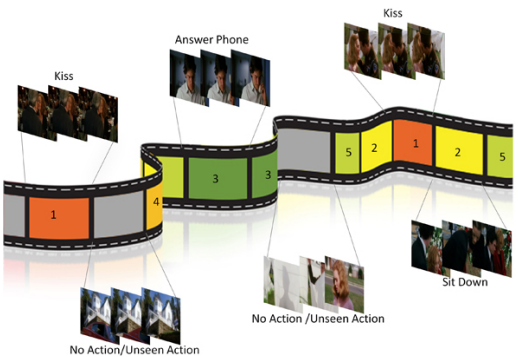
3. http ://www.ad-exchange.fr/adform-une-plateforme-integree-au-maximum-y-compris-pour-lutter-contre-la-fraude-18634/

Applications

Web

Classification automatique de documents par mots-clés (textes, images, sons, vidéos...)

4



5

4. http ://www.analyticbridge.com/profiles/blogs/text-analysis-101-a-basic-understanding-for-business-users
5. http ://www.ccs.neu.edu/home/eelhami/research.htm

Conclustions et perspectives

Conclusions

- Le clustering de Dirichlet peut être appliqué à plusieurs domaines surtout quand on connait pas le nombre de clusters à choisir.
- Nouvelles notions mathématiques.
- Programmation sous R et python.
- Formation en lateX.

Perspectives

- Utiliser les métriques pour voir la fiabilité de l'algorithme.
- Mettre des lois sur le paramètre α et voir le comportement du nombre des clusters.