Projet M2 IM

-

Introduction aux Topic Models

Ulysse Herbach

17 décembre 2014

Introduction

On dispose d'une grande quantité de textes sous forme d'un corpus.

- **Problème :** on voudrait être capable d'analyser ces textes de manière *automatique* pour en tirer l'information la plus pertinente possible.
- **Idée :** comprendre la structure sous-jacente du corpus de textes en dégageant des thèmes, ou *topics*, supposés présents dans le corpus mais non connus à l'avance.
- **Objectif :** fournir des informations plus fines qu'un simple résumé du corpus (ex : polysémie).

Plan

1. Les principaux Topic Models
2. Cas de pLSI : estimation des paramètres
3. Application à une vraie base de données
4. Conclusion et perspectives

Les principaux Topic Models

Hypothèses simplificatrices générales :

- *bag-of-words* : l'ordre des mots n'a pas d'importance sur la façon de générer un document.
- *bag-of-documents* : les documents n'ont pas d'ordre spécifique dans le corpus.

Les principaux Topic Models

Hypothèses simplificatrices générales :

- *bag-of-words* : l'ordre des mots n'a pas d'importance sur la façon de générer un document.
- *bag-of-documents* : les documents n'ont pas d'ordre spécifique dans le corpus.

Cadre probabiliste :

- L'idée générale est de décrire la loi d'un corpus aléatoire
- Les thèmes sont des variables aléatoires *latentes* (i.e. non observées)
- On fixe un dictionnaire dans lequel seront tirés tous les mots du corpus, on note $M \in \mathbb{N}^*$ sa taille.
- Pour générer un document constitué de N mots, on considère des variables aléatoires $W_1, \ldots, W_N$, à valeurs dans $\{1, \ldots, M\}$, et on assimile le document à la variable aléatoire $D = (W_1, \ldots, W_N)$.

Notations

- Pour $(w_1, \ldots, w_N) \in \{1, \ldots, M\}^N$, la probabilité qu'un document D soit constitué des mots $(w_1, \ldots, w_N)$ est notée $P_D(w_1, \ldots, w_N)$.
- Quand il n'y aura pas de confusion possible, on notera, pour toute variable aléatoire discrète X et pour tout x dans l'image de X :

$$p(x) = P_X(x) = \mathbb{P}(X = x).$$

- Si Y est une autre v.a. et si y est dans l'image de Y avec $p(y) > 0$, on notera $p(x|y) = \mathbb{P}(X = x|Y = y)$.

Le modèle *unigram*

C'est le modèle probabiliste le plus simple possible. On suppose que :

- les documents sont générés indépendamment et de la même façon ;
- les mots de chaque document sont générés de manière indépendante, à partir d'une même mesure de probabilité discrète.

La distribution associée est donc :

$$P_D(w_1, \ldots, w_N) = \prod_{n=1}^{N} p(w_n)$$

Une façon de générer un corpus de m documents est de fixer N et M, puis considérer des variables i.i.d. $D_1, \ldots, D_m$.
Les documents contiennent alors tous le même nombre de mots.

Le modèle *mixture of unigrams*

On introduit une variable latente "topic" notée Z, à valeurs dans $\{1, \ldots, K\}$ où K est le nombre de topics. On fait les hypothèses suivantes :

- les documents sont générés de manière i.i.d. ;
- chaque document est associé à un unique topic ;
- les mots d'un document sont générés indépendamment, à partir d'une même proba discrète dépendant du topic associé au document.

La distribution associée est donc :

$$P_D(w_1, \ldots, w_N) = \sum_{z=1}^{K} p(z) \prod_{n=1}^{N} p(w_n|z)$$

Le modèle *probabilistic Latent Semantic Indexing* (pLSI)

Ici, on fixe dès le départ le nombre N de documents générés.
Chaque mot du corpus correspond à la réalisation d'une v.a. (D, W, Z) :

- $D \in \{1, \ldots, N\}$ est l'indice du document dans lequel se trouve le mot ;
- $W \in \{1, \ldots, M\}$ est l'indice du mot dans le dictionnaire ;
- $Z \in \{1, \ldots, K\}$ est le topic sous-jacent à *cette occurrence* du mot.

Le modèle *probabilistic Latent Semantic Indexing* (pLSI)

Ici, on fixe dès le départ le nombre N de documents générés.
Chaque mot du corpus correspond à la réalisation d'une v.a. (D, W, Z) :

- $D \in \{1, \dots, N\}$ est l'indice du document dans lequel se trouve le mot ;
- $W \in \{1, \dots, M\}$ est l'indice du mot dans le dictionnaire ;
- $Z \in \{1, \dots, K\}$ est le topic sous-jacent à *cette occurrence* du mot.

On peut décrire le modèle sous la forme d'une famille de n variables $((D_i, W_i, Z_i))_{1 \leqslant i \leqslant n}$, où n est le nombre total de mots du corpus.

On fait alors les hypothèses suivantes :

- les variables (D_i, W_i, Z_i), $1 \leqslant i \leqslant n$ sont i.i.d ;
- $\forall i \in \{1, \dots, n\}, \quad \mathcal{L}(D_i, W_i \mid Z_i) = \mathcal{L}(D_i \mid Z_i) \otimes \mathcal{L}(W_i \mid Z_i)$.

Au final, la distribution d'une observation (d, w) est

$$p(d, w) = p(d) \sum_{z=1}^{K} p(w|z) p(z|d).$$

Le modèle *Latent Dirichlet Allocation* (LDA)

But : nombre de paramètres indépendant du nombre de documents, en gardant la souplesse des topics. La clé est un aléa supplémentaire...

On prend Z de loi multinomiale *aléatoire*, générée par une loi de *Dirichlet* de paramètre $\alpha = (\alpha_1, \ldots, \alpha_K)$ fixé, définie par sa densité sur $\mathbb{R}^K$:

$$f_\alpha(\theta) = f_\alpha(\theta_1, \ldots, \theta_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1}$$

si $\theta \in \Delta_K$ et $f_\alpha(\theta) = 0$ sinon, où
$\Delta_K = \left\{ \theta \in \mathbb{R}^K \mid \forall i \in \{1, \ldots, K\},\ \theta_i > 0 \quad \text{et} \quad \sum_{i=1}^K \theta_i = 1 \right\}$.

Le modèle *Latent Dirichlet Allocation* (LDA)

But : nombre de paramètres indépendant du nombre de documents, en gardant la souplesse des topics. La clé est un aléa supplémentaire...

On prend Z de loi multinomiale *aléatoire*, générée par une loi de *Dirichlet* de paramètre $\alpha = (\alpha_1, \ldots, \alpha_K)$ fixé, définie par sa densité sur $\mathbb{R}^K$:

$$f_\alpha(\theta) = f_\alpha(\theta_1, \ldots, \theta_K) = \frac{\Gamma\left(\sum_{i=1}^K \alpha_i\right)}{\prod_{i=1}^K \Gamma(\alpha_i)} \theta_1^{\alpha_1 - 1} \ldots \theta_K^{\alpha_K - 1}$$

si $\theta \in \Delta_K$ et $f_\alpha(\theta) = 0$ sinon, où
$\Delta_K = \left\{\theta \in \mathbb{R}^K \mid \forall i \in \{1, \ldots, K\},\ \theta_i > 0 \quad \text{et} \quad \sum_{i=1}^K \theta_i = 1\right\}$.

Pour générer $D = (W_1, \ldots, W_N)$, on procède de la manière suivante :

1. on tire θ selon une loi $\mathrm{Dir}(\alpha)$;
2. pour tout $n \in \{1, \ldots, N\}$:
   1. on tire un topic z_n selon une loi multinomiale $\mathcal{M}_K(1, \theta)$;
   2. on tire un mot w_n dans $\{1, \ldots, M\}$ avec une probabilité $p(w_n | z_n)$ fixée.

Le modèle *Latent Dirichlet Allocation* (LDA)

Au final, pour un document donné, les v.a. en jeu sont Θ, $Z_1, \ldots, Z_N$ et $W_1, \ldots, W_N$ avec les hypothèses suivantes :

- $\Theta \sim \mathrm{Dir}(\alpha)$
- $\forall n \in \{1, \ldots, N\}$, $\mathcal{L}(Z_n|\Theta) = \mathcal{M}_K(1, \Theta)$ et $\mathbb{P}(W_n = j | Z_n = e_i) = p_\beta(j, e_i) = \beta_{i,j}$ fixé,

pour $(i, j) \in \{1, \ldots, K\} \times \{1, \ldots, M\}$. Le nombre total de paramètres du modèle est donc $K + KM$.

La distribution associée au modèle est :

$$P_D(w_1, \ldots, w_N) = \int_{\Delta_K} \left(\prod_{n=1}^{N} p(w_n|\theta) \right) f_\alpha(\theta)\, d\theta \tag{1}$$

avec

$$p(w_n|\theta) = \sum_{k=1}^{K} p_\beta(w_n|z_k) p(z_k|\theta).$$

Cas de pLSI : estimation des paramètres

On doit estimer $K + KN + KM$ paramètres, notés

$$\alpha(z) = p(z), \quad \beta(d, z) = p(d|z) \quad \text{et} \quad \gamma(w, z) = p(w|z).$$

On considère des "observations" i.i.d $(D_i, W_i, Z_i)_{1 \leqslant i \leqslant n}$ (mais on n'observe pas les Z_i). Estimation par maximum de la log-vraisemblance *observée* :

$$\ell(\bar{d}, \bar{w}; \alpha, \beta, \gamma) = \sum_{d=1}^{N} \sum_{w=1}^{M} \bar{n}(d, w) \ln \left(\sum_{z=1}^{K} \alpha(z) \beta(d, z) \gamma(w, z) \right)$$

où $\bar{d} = (d_1, \ldots, d_n)$ et $\bar{w} = (w_1, \ldots, w_n)$ et $\bar{n}(d, w)$ est le nombre d'observations (d_i, w_i) égales à (d, w). Difficile de la maximiser...

Cas de pLSI : estimation des paramètres

On doit estimer $K + KN + KM$ paramètres, notés

$$\alpha(z) = p(z), \quad \beta(d, z) = p(d|z) \quad \text{et} \quad \gamma(w, z) = p(w|z).$$

On considère des "observations" i.i.d $(D_i, W_i, Z_i)_{1 \leqslant i \leqslant n}$ (mais on n'observe pas les Z_i). Estimation par maximum de la log-vraisemblance *observée* :

$$\ell(\bar{d}, \bar{w}; \alpha, \beta, \gamma) = \sum_{d=1}^{N} \sum_{w=1}^{M} \bar{n}(d, w) \ln \left(\sum_{z=1}^{K} \alpha(z) \beta(d, z) \gamma(w, z) \right)$$

où $\bar{d} = (d_1, \ldots, d_n)$ et $\bar{w} = (w_1, \ldots, w_n)$ et $\bar{n}(d, w)$ est le nombre d'observations (d_i, w_i) égales à (d, w). Difficile de la maximiser...
On se tourne vers la log-vraisemblance "complétée" :

$$\ell_c(\bar{d}, \bar{w}, \bar{z}; \alpha, \beta, \gamma) = \sum_{i=1}^{n} \ln \alpha(z_i) + \ln \beta(d_i, z_i) + \ln \gamma(w_i, z_i).$$

Le calcul de celle-ci est impossible sans observer les Z_i, mais il est possible d'obtenir une estimation d'un max local de ℓ via ℓ_c : c'est l'algorithme EM.

Algorithme EM

On note $\theta = (\alpha, \beta, \gamma)$. Etant donné un paramètre initial $\theta^{(0)}$, on veut converger vers un certain $\theta^{(\infty)}$ qui réalise un maximum local de ℓ.

Expectation : on calcule la fonction

$$\theta \mapsto Q(\theta, \theta^{(t)}) = \mathbb{E}_{\theta^{(t)}} \left[\ell_c(\bar{D}, \bar{W}, \bar{Z}; \theta) | \bar{D}, \bar{W} \right].$$

On a alors :

$$Q(\theta, \theta^{(t)}) = \sum_{d,w,z} \bar{n}(d,w) p^{(t)}(z|d,w)(\ln \alpha(z) + \ln \beta(d,z) + \ln \gamma(w,z))$$

où

$$p^{(t)}(z|d,w) = \frac{\alpha^{(t)}(z)\beta^{(t)}(d,z)\gamma^{(t)}(w,z)}{\sum_{z'} \alpha^{(t)}(z')\beta^{(t)}(d,z')\gamma^{(t)}(w,z')}.$$

Algorithme EM

Maximization : on calcule le point $\theta^{(t+1)}$ qui réalise le max de $Q(\cdot, \theta^{(t)})$. Grâce au théorème des extrema liés (contraintes associées aux probabilités), on trouve :

$$\alpha^{(t+1)}(z) = \frac{1}{n} \sum_{d,w} \bar{n}(d,w) p^{(t)}(z|d,w)$$

$$\beta^{(t+1)}(d,z) = \frac{\sum_w \bar{n}(d,w) x p^{(t)}(z|d,w)}{\sum_{d',w} \bar{n}(d',w) p^{(t)}(z|d',w)}$$

$$\gamma^{(t+1)}(w,z) = \frac{\sum_d \bar{n}(d,w) p^{(t)}(z|d,w)}{\sum_{d,w'} \bar{n}(d,w') p^{(t)}(z|d,w')}$$

avec $n = \sum_{d,w} \bar{n}(d,w)$, le nombre total d'observations.

Croissance de la vraisemblance

L'algorithme EM ainsi construit fournit une suite de paramètres $(\theta^{(t)})_{t\in\mathbb{N}}$. L'intuition que l'on avait en prenant l'espérance conditionnelle est justifiée par le résultat suivant :

Théorème

La vraisemblance est croissante le long de l'algorithme EM. En d'autres termes, la suite $(\theta^{(t)})_{t\in\mathbb{N}}$ vérifie

$$\forall t \in \mathbb{N}, \quad \ell(\bar{d}, \bar{w}; \theta^{(t+1)}) \geq \ell(\bar{d}, \bar{w}; \theta^{(t)}).$$

De plus, si $\theta^{(t)}$ n'est pas un point stationnaire, i.e. n'annule pas le gradient de ℓ, alors

$$\ell(\bar{d}, \bar{w}; \theta^{(t+1)}) > \ell(\bar{d}, \bar{w}; \theta^{(t)}).$$

Remarque : A priori on est condamné à ne converger que vers un point selle ou un maximum local dépendant du point initial $\theta^{(0)}$.

Application de pLSI à la base **20 Newsgroups**

Environ 20000 messages dans 20 catégories différentes :

Indice	Intitulé	Indice	Intitulé
1	« AltAtheism »	11	« Hockey »
2	« CompG »	12	« SciCrypt »
3	« CompOS »	13	« SciElectronics »
4	« CompIBM »	14	« SciMed »
5	« CompMac »	15	« SciSpace »
6	« CompWindows »	16	« SocReligion »
7	« MiscForsale »	17	« TalkGuns »
8	« Autos »	18	« TalkMideast »
9	« Motos »	19	« TalkMisc »
10	« Baseball »	20	« TalkReligion »

Nettoyage des données : On utilise un dictionnaire de *stopwords* (caractères alphanumériques, conjonctions de coordination, ponctuation...)
On ne garde que les mots qui apparaissent au moins 2 fois dans le corpus (fautes de frappes, néologismes personnels...)

Application de pLSI à la base **20 Newsgroups**

Constantes du modèle : $K = 20$, $M = 59809$ et $N = 19997$.

Application de pLSI à la base **20 Newsgroups**

Constantes du modèle : $K = 20$, $M = 59809$ et $N = 19997$.

Structure *sparse* : on ne considère que les *entrées utiles* : $\bar{n}(d, w) \neq 0$.
On passe de $n = 2082550$ observations à $u = 1501986$ entrées utiles.

Application de pLSI à la base **20 Newsgroups**

Constantes du modèle : $K = 20$, $M = 59809$ et $N = 19997$.

Structure *sparse* : on ne considère que les *entrées utiles* : $\bar{n}(d, w) \neq 0$.
On passe de $n = 2082550$ observations à $u = 1501986$ entrées utiles.

L'algorithme EM estime $p(z), p(d|z)$ et $p(w|z)$. Pour étudier l'efficacité, on considère qu'un document a toujours un topic "principal".

Clustering : À un document d donné, on associe le topic z ssi

$$p(z|d) > p(z'|d), \quad \forall z' \neq z.$$

On se ramène à maximiser $p(z)p(d|z)$, grâce à la formule de Bayes

$$p(z|d) = \frac{p(z)}{p(d)} p(d|z).$$

Mesure d'efficacité

On a besoin d'identifier les topics z (ou *clusters*) et les "vraies" catégories.
On attribue à chaque cluster la catégorie majoritairement présente.
Soit i un cluster et soit j la catégorie à laquelle il a été associé.

- $C(i,j)$ = nombre de documents de la catégorie j affectés au cluster i
- $T_{clu}(i)$ = nombre de documents du cluster i
- $T_{cat}(j)$ = nombre de documents de la catégorie j

Mesure d'efficacité

L'efficacité du cluster i est $F_i = \frac{2C(i,j)}{T_{clu}(i) + T_{cat}(j)}, \forall i \in \{1, \ldots, K\}.$

Mesure d'efficacité

On a besoin d'identifier les topics z (ou *clusters*) et les "vraies" catégories.
On attribue à chaque cluster la catégorie majoritairement présente.
Soit i un cluster et soit j la catégorie à laquelle il a été associé.

- $C(i,j)$ = nombre de documents de la catégorie j affectés au cluster i
- $T_{clu}(i)$ = nombre de documents du cluster i
- $T_{cat}(j)$ = nombre de documents de la catégorie j

Mesure d'efficacité

L'efficacité du cluster i est $F_i = \dfrac{2C(i,j)}{T_{clu}(i) + T_{cat}(j)}, \forall i \in \{1, \ldots, K\}.$

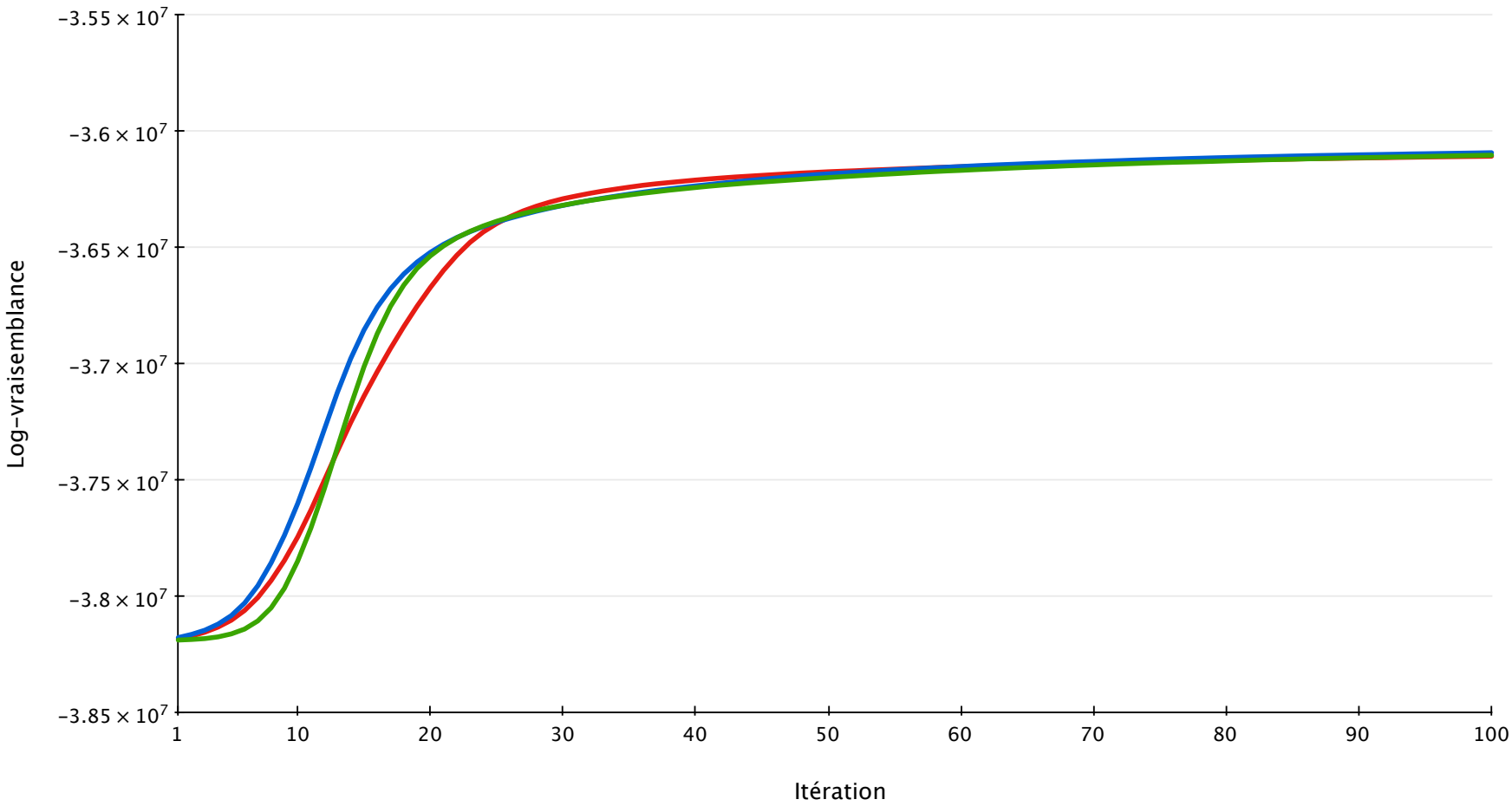
Cohérence de la mesure :

$$0 \leqslant C(i,j) \leqslant \min(T_{clu}(i), T_{cat}(j)) \leqslant \frac{1}{2}(T_{clu}(i) + T_{cat}(j))$$

d'où $0 \leqslant F_i \leqslant 1$, avec $F_i = 1$ ssi $C(i,j) = T_{clu}(i) = T_{cat}(j)$.

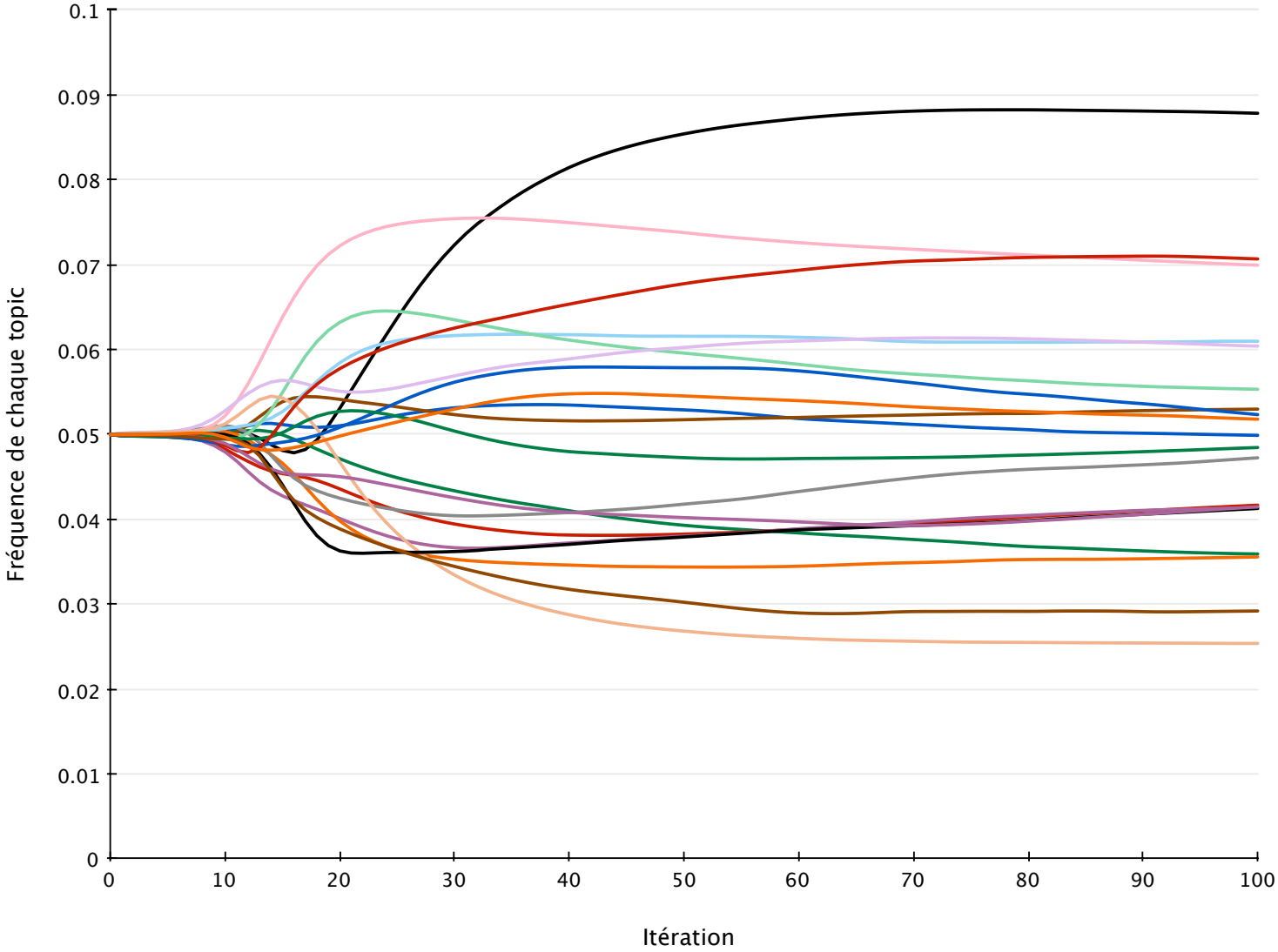
Résultats de algorithme EM

Log-vraisemblance des observations sur 100 itérations



Résultats de algorithme EM

Évolution des probabilités des 20 topics le long de l'algorithme EM



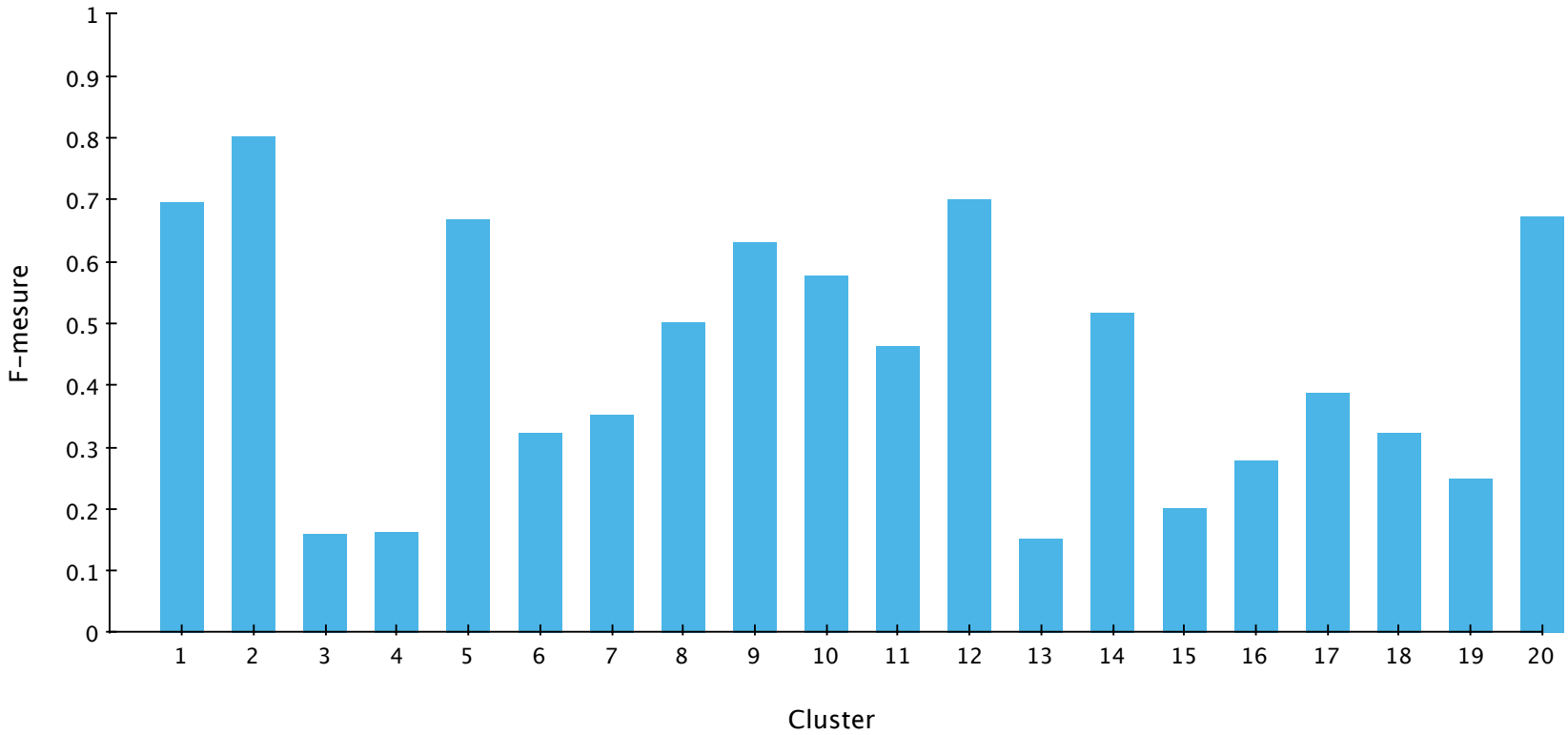
Résultats de algorithme EM

Clustering : quelques thèmes générés par le modèle pLSI

« Médecine » 5 (SciMed)	« Baseball » 9 (Baseball)	« Espace » 10 (SciSpace)	« Informatique » 14 (CompG)
medicine	braves	shuttle	file
risk	sox	payload	format
candida	pitcher	radar	pub
diseases	hitter	comet	information
women	cubs	titan	send
physician	batting	jet	number
hicnet	alomar	orbiter	address
infections	clemens	aerospace	contact
physicians	phillies	centaur	ftp
lyme	innings	baalke	site

Résultats de algorithme EM

Efficacité de chacun des 20 clusters en sortie de l'algorithme EM



Conclusion et perspectives

- Le modèle pLSI semble avoir très bien capté les thèmes sous-jacents de la base de données. Temps de calcul acceptable.
- Parfois, un topic capte autre chose qu'une catégorie existante.
- les Topic Models peuvent potentiellement s'appliquer à n'importe quel type de données (génomique, images...).
- Il serait intéressant de poursuivre l'étude en implémentant LDA (algorithme "EM variationnel", échantillonneur de Gibbs).
- Perfectionnement des modèles dans deux directions différentes :
  - modèles plus souples (moins d'hypothèses) ;
  - modèles spécialisés (hypothèses spécifiques).

Références

E. Bart, M. Welling, and P. Perona, *Unsupervised organization of image collections : Taxonomies and beyond*, Trans. Pattern Recognit. Mach. Intell. **33** (2010).

David M. Blei, Andrew Y. Ng, and Michael I. Jordan, *Latent Dirichlet Allocation*, Journal of Machine Learning Research (2003), no. 3, 993–1022.

Y.S. Chow and H Teicher, *Probability theory : independence, interchangeability, martingales*, Springer Verlag, 1997.

William M. Darling, *A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling*, Tech. report, School of Computer Science, University of Guelph, 2011.

S. Gaïffas, *Mélanges de lois et algorithme EM*, Cours donné à l'École Polytechnique, 2014.

Thomas Hofmann, *Probabilistic Latent Semantic Indexing*, Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrievial (1999).

J. Pritchard, M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data*, Genetics **155** (2000).

Bonus : preuve de l'étape *Maximization* de EM

Théorème (Extrema liés)

Soient $p, r \in \mathbb{N}^*$, U un ouvert de $\mathbb{R}^p$, et $f, g_1, \ldots, g_r \in \mathcal{C}^1(U, \mathbb{R})$. On considère l'ensemble suivant :

$$\Gamma = \{x \in U \mid \forall i \in \{1, \ldots, r\},\ g_i(x) = 0\}.$$

On suppose que $f_{|\Gamma}$ admet un extremum local en $a \in \Gamma$ et que la famille $(\mathrm{d}g_i(a))_{1 \leq i \leq r}$ est libre dans $(\mathbb{R}^p)'$. Alors les formes linéaires $\mathrm{d}f(a), \mathrm{d}g_1(a), \ldots, \mathrm{d}g_r(a)$ sont liées. En d'autres termes, il existe $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$ appelés multiplicateurs de Lagrange, tels que

$$\mathrm{d}f(a) = \sum_{i=1}^{r} \lambda_i \, \mathrm{d}g_i(a).$$

On note $\theta = (\alpha, \beta, \gamma) \in \mathbb{R}^K \times \mathbb{R}^{KN} \times \mathbb{R}^{KM} = \mathbb{R}^{K+KN+KM}$. On veut appliquer le théorème avec $p = K + KN + KM$ et $f : \theta \mapsto Q(\theta, \theta^{(t)})$.

Bonus : preuve de l'étape *Maximization* de EM

Les $r = 1 + 2K$ contraintes de maximisation sont :

$$\sum_{z=1}^{K} \alpha(z) = 1 \quad \text{et} \quad \forall z \in \{1 \dots, K\}, \quad \sum_{d=1}^{N} \beta(d, z) = \sum_{w=1}^{M} \gamma(w, z) = 1$$

On peut les réécrire sous la forme

$$g_\alpha(\theta) = 0 \quad \text{et} \quad \forall z \in \{1 \dots, K\}, \quad g_{\beta,z}(\theta) = g_{\gamma,z}(\theta) = 0$$

où $g_\alpha(\theta) = \sum_{z=1}^{K} \alpha(z) - 1$, $g_{\beta,z}(\theta) = \sum_{d=1}^{N} \beta(d, z) - 1$ et $g_{\gamma,z}(\theta) = \sum_{w=1}^{M} \gamma(w, z) - 1$. D'après le théorème, il existe $2K + 1$ réels λ_α, $(\lambda_{\beta,z})_{1 \leqslant z \leqslant K}$ et $(\lambda_{\beta,z})_{1 \leqslant z \leqslant K}$ tels que

$$\mathrm{d}f(\theta^{(t+1)}) = \lambda_\alpha \, \mathrm{d}g_\alpha(\theta^{(t+1)}) + \sum_{z=1}^{K} \lambda_{\beta,z} \, \mathrm{d}g_{\beta,z}(\theta^{(t+1)}) + \sum_{z=1}^{K} \lambda_{\gamma,z} \, \mathrm{d}g_{\gamma,z}(\theta^{(t+1)}).$$

Bonus : preuve de l'étape *Maximization* de EM

Ceci se traduit par : $\forall d \in \{1 \dots, N\}$, $\forall w \in \{1 \dots, M\}$, $\forall z \in \{1 \dots, K\}$,

$$\frac{1}{\alpha^{(t+1)}(z)} \sum_{d=1}^{N} \sum_{w=1}^{M} \bar{n}(d,w) p^{(t)}(z|d,w) = \lambda_{\alpha},$$

$$\frac{1}{\beta^{(t+1)}(d,z)} \sum_{w=1}^{M} \bar{n}(d,w) p^{(t)}(z|d,w) = \lambda_{\beta,z}$$

$$\frac{1}{\gamma^{(t+1)}(w,z)} \sum_{d=1}^{N} \bar{n}(d,w) p^{(t)}(z|d,w) = \lambda_{\gamma,z}$$

On conclut en utilisant le fait que $\theta^{(t+1)} \in \Gamma$.

Bonus : preuve de la croissance de la vraisemblance

Dans ce qui suit, on note pour simplifier $\bar{X} = (\bar{D}, \bar{W})$, $\bar{x} = (\bar{d}, \bar{w})$, $\mathbb{P}_\theta(\bar{X}) = \mathbb{P}_\theta(\bar{X} = \bar{x})$ et $\mathbb{P}_\theta(\bar{Z}|\bar{X}) = \mathbb{P}_\theta(\bar{Z} = \bar{z}, \bar{X} = \bar{x})$. L'idée est de conditionner par la variable latente $\bar{Z}$ en utilisant la relation :

$$\mathbb{P}_\theta(\bar{Z}|\bar{X}) = \frac{\mathbb{P}_\theta(\bar{Z}, \bar{X})}{\mathbb{P}_\theta(\bar{X})}.$$

On a alors, pour tout $\theta \in U$:

$$\ell(\bar{X}; \theta) = \ln\left(\mathbb{P}_\theta(\bar{Z}, \bar{X}) \frac{\mathbb{P}_\theta(\bar{X})}{\mathbb{P}_\theta(\bar{Z}, \bar{X})}\right) = \ln\left(\mathbb{P}_\theta(\bar{Z}, \bar{X})\right) - \ln\left(\mathbb{P}_\theta(\bar{Z}|\bar{X})\right),$$

ce qui donne, en notant $\ell(\bar{Z}|\bar{X}; \theta) = \ln\left(\mathbb{P}_\theta(\bar{Z}|\bar{X})\right)$:

$$\ell(\bar{X}; \theta) = \ell_c(\bar{X}, \bar{Z}; \theta) - \ell(\bar{Z}|\bar{X}; \theta).$$

Bonus : preuve de la croissance de la vraisemblance

Soit $t \in \mathbb{N}$, on considère que $\theta^{(t)}$ et $\theta^{(t+1)}$ sont fixés. On a alors, pour tout $\theta \in U$, en prenant l'espérance par rapport à la loi $\mathbb{P}_{\theta^{(t)}}(\cdot|\bar{X})$:

$$\begin{aligned}\ell(\bar{X};\theta) = \mathbb{E}_{\theta^{(t)}}[\ell(\bar{X};\theta)|\bar{X}] &= \mathbb{E}_{\theta^{(t)}}[\ell_c(\bar{X},\bar{Z};\theta)|\bar{X}] - \mathbb{E}_{\theta^{(t)}}[\ell(\bar{Z}|\bar{X};\theta)|\bar{X}] \\ &= Q(\theta,\theta^{(t)}) - Q_1(\theta,\theta^{(t)})\end{aligned}$$

où $Q_1(\theta,\theta^{(t)}) = \mathbb{E}_{\theta^{(t)}}[\ell(\bar{Z}|\bar{X};\theta)|\bar{X}]$. On a finalement :

$$\begin{aligned}\ell(\bar{X};\theta^{(t+1)}) - \ell(\bar{X};\theta^{(t)}) = \underbrace{Q(\theta^{(t+1)},\theta^{(t)})}_{①} - Q(\theta^{(t)},\theta^{(t)}) \\ - (\underbrace{Q_1(\theta^{(t+1)},\theta^{(t)})}_{②} - Q_1(\theta^{(t)},\theta^{(t)}))\end{aligned}$$

Tout d'abord, il est clair que ① $\geqslant 0$ par définition de $\theta^{(t+1)}$.

Bonus : preuve de la croissance de la vraisemblance

Montrons que ② $\leqslant 0$. Pour tout $\theta \in U$, on a par Jensen :

$$\begin{aligned} Q_1(\theta, \theta^{(t)}) - Q_1(\theta^{(t)}, \theta^{(t)}) &= \mathbb{E}_{\theta^{(t)}} \left[\ln \left(\frac{\mathbb{P}_\theta(\bar{Z}|\bar{X})}{\mathbb{P}_{\theta^{(t)}}(\bar{Z}|\bar{X})} \right) |\bar{X} \right] \\ &\leqslant \ln \left(\mathbb{E}_{\theta^{(t)}} \left[\frac{\mathbb{P}_\theta(\bar{Z}|\bar{X})}{\mathbb{P}_{\theta^{(t)}}(\bar{Z}|\bar{X})} |\bar{X} \right] \right) \\ &\leqslant \ln \left(\sum_{\bar{z}} \mathbb{P}_\theta(\bar{Z} = \bar{z}|\bar{X}) \right) \leqslant \ln(1) = 0 \end{aligned}$$

et on conclut en appliquant ceci en $\theta = \theta^{(t+1)}$. Si $\nabla_\theta \ell(\bar{d}, \bar{w}; \theta^{(t)}) \neq 0$, l'inégalité précédente montre que la fonction $\theta \mapsto Q_1(\theta, \theta^{(t)})$ est maximale en $\theta = \theta^{(t)}$. On en déduit que $\nabla_\theta Q_1(\theta, \theta^{(t)}) = 0$ en $\theta = \theta^{(t)}$, puis que

$$\nabla_\theta Q(\theta, \theta^{(t)}) = \nabla_\theta \ell(\bar{d}, \bar{w}; \theta) \neq 0$$

en $\theta = \theta^{(t)}$, d'où $Q(\theta^{(t+1)}, \theta^{(t)}) > Q(\theta^{(t)}, \theta^{(t)})$ et $\ell(\bar{X}; \theta^{(t+1)}) > \ell(\bar{X}; \theta^{(t)})$.