Modèles à variables latentes pour l'écologie et la biologie

Examen 2024

Le *Latent Dirichlet Allocation (LDA) model* est un modèle à variables latentes permettant d'analyser des textes, et notamment de trouver de façon non-supervisée des *topics* (sujets) abordés dans les textes analysés, ces topics étant décrits par des assemblages de mots.

Supposons qu'on cherche à analyser D documents. Chaque document d ($d = 1, \dots, D$) est composé de n_d mots. En général, les textes ont été prétraités, on a enlevé les mots de liaison qui n'apportent pas de sens sur le thème abordé, enlevé les marques de pluriel, conjuguaison, etc...

On suppose que les n_d mots appartiennent à un dictionnaire de taille V. Soit Y_{di} la variable représentant le i-eme du d-ème document.

$$Y_{di} = v \Leftrightarrow \text{le } i\text{-ème mot du document } d \text{ est le } v\text{-ème mot du dictionnaire}$$

Exemple

Texte 1 :	réchauffement	-	carbone	-	pollution	-	innondation	-	dégat
Texte 2 :	réchauffement	-	tempête	-	arbre	-			
Texte 3 :	chomage	-	pauvreté	-	inflation	-	logement		
Texte 4 :	logement	-	passoire	-	thermique	-	aide	-	isolation

Le dictionnaire est composé de 15 mots

$$\{\text{réchauffement}, \text{carbone}, \text{pollution}, \text{innondation}, \text{dégat}, \text{tempête}, \text{arbre}, \text{chomage}, \dots\}$$

et on obtient $Y_{11} = 1, Y_{12} = 2$, pour le document $d = 1$ et $Y_{21} = 1, Y_{22} = 6$ pour le document $d = 2$.

L'idée du modèle LDA est que les mots utilisés dans chaque texte dépendent des topics qui sont abordés dans le texte et les topics sont caractérisés par des fréquences de mots du dictionnaire.

Le modèle LDA Soit K le nombre de topics. A chaque topic k, on associe un vecteur de fréquences des mots du dictionnaire:

$$\boldsymbol{\phi}_k = (\phi_{k1}, \dots, \phi_{kV}) \in [0,1]^V \quad \text{tel que} \quad \sum_{v=1}^{V} \phi_{kv} = 1, \qquad \forall k = 1, \dots, K.$$

Pour chaque texte d, on suppose que plusieurs topics sont abordés en proportion propre au texte d. On note $\boldsymbol{\eta}_d = (\eta_{d1}, \dots, \eta_{dK}) \in [0,1]^K$ le vecteur des "topic proportions" du texte d avec $\sum_{k=1}^{K} \eta_{dk} = 1$.

Pour chaque mot i du texte d, on note Z_{di} le topic qui lui est associé $Z_{di} \in \{1, \dots, K\}$. Les variables Z_{di} sont indépendantes et identiquement distribuées et

$$P(Z_{di} = k) = \eta_{dk}.$$

Ensuite, la loi de Y_{di} (i.e. le mot i du texte d) ne dépend que du topic auquel il est rattaché:

$$P(Y_{di} = v | Z_{di} = k) = \phi_{kv}.$$

Notations On introduit les notations suivantes:

$$\begin{array}{lcllcl}
\boldsymbol{Y}_d & := & (Y_{di})_{i=1,\dots,n_d} & \boldsymbol{Y} & := & (\boldsymbol{Y}_d)_{d=1,\dots,D} \\
\boldsymbol{Z}_d & := & (Z_{di})_{i=1,\dots,n_d} & \boldsymbol{Z} & := & (\boldsymbol{Z}_d)_{d=1,\dots,D} \\
\boldsymbol{\phi}_k & := & (\phi_{kv})_{v=1,\dots V} & \boldsymbol{\phi} & := & (\boldsymbol{\phi}_k)_{k=1,\dots K} \\
\boldsymbol{\eta}_d & := & (\eta_{dk})_{k=1,\dots K} & \boldsymbol{\eta} & := & (\boldsymbol{\eta}_d)_{d=1,\dots D} \\
\theta & := & (\boldsymbol{\phi}, \boldsymbol{\eta}) & & & \\
Y_{div} & := & \mathbb{1}_{Y_{di}=v}, & \text{donc} & & \sum_{v=1}^{V} Y_{div} = 1 \\
Z_{dik} & := & \mathbb{1}_{Z_{di}=k}, & \text{donc} & & \sum_{k=1}^{K} Z_{dik} = 1
\end{array}$$

Inférence On travaille de façon non-supervisée. Le nombre de topics K et la caractérisation des topics $(\phi_{kv})_{k=1,\dots,K,v=1,\dots,V}$ ne sont pas connus. Les Z sont des variables latentes. On cherche à estimer θ à K connu.

Partie 1. Inférence fréquentiste (50%)

1. Montrer que la log vraisemblance complète $\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z})$ s'écrit:

$$\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) = \sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}\right]\log\phi_{kv} + \sum_{d=1}^{D}\sum_{k=1}^{K}\left[\sum_{i=1}^{n_d} Z_{dik}\right]\log\eta_{dk}$$

$$\begin{aligned}
\log p_\theta(\boldsymbol{Y}, \boldsymbol{Z}) &= \sum_{d=1}^{D}\sum_{i=1}^{n_d} \log p(Y_{di}|Z_{di}) + \log p_\theta(Z_{di}) \\
&= \sum_{d=1}^{D}\sum_{i=1}^{n_d}\sum_{k=1}^{K} \log p(Y_{di}|Z_{di}=k) + \sum_{d=1}^{D}\sum_{i=1}^{n_d}\sum_{k=1}^{K} \log P_\theta(Z_{di}=k) \\
&= \sum_{d=1}^{D}\sum_{i=1}^{n_d}\sum_{k=1}^{K}\sum_{v=1}^{V} Z_{dik}Y_{div}\log\phi_{kv} + \sum_{d=1}^{D}\sum_{i=1}^{n_d}\sum_{k=1}^{K} Z_{dik}\log\eta_{dk} \\
&= \sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}\right]\log\phi_{kv} + \sum_{d=1}^{D}\sum_{k=1}^{K}\left[\sum_{i=1}^{n_d} Z_{dik}\right]\log\eta_{dk}
\end{aligned}$$

2. Donner les estimateurs du maximum de vraisemblance de θ si on observe à la fois les textes $\boldsymbol{Y}$ et les topics $\boldsymbol{Z}$. Interpréter les statistiques obtenues.

- *A propos de* ϕ_{kv}

$$\frac{\partial}{\partial \phi_{kv}}\left[\log p_\theta(\boldsymbol{Y},\boldsymbol{Z}) - \lambda_k\left(\sum_{v=1}^{V}\phi_{kv}-1\right)\right] = \left[\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}\right]\frac{1}{\phi_{kv}} - \lambda_k \quad (1)$$

$$\widehat{\phi}_{kv} = \frac{\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}}{\lambda_k} \quad (2)$$

Sous la contrainte $\sum_{v=1}^{V}\phi_{kv}=1$, on obtient

$$\lambda_k = \sum_{v=1}^{V}\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div} = \sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}\underbrace{\sum_{v=1}^{V} Y_{div}}_{=1} = \sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik} = N_k$$

N_k est le nombre de mots appartenant au topic k dans toute la collection de textes.

$$\widehat{\phi}_{kv} = \frac{\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}}{N_k}$$

C'est le nombre de fois où on a vu le mot v associé au topic k divisé par le nombre de mots appartenant au topic k.

- *A propos de* η_{dk}

$$\frac{\partial}{\partial \eta_{dk}}\left[\log p_\theta(\boldsymbol{Y},\boldsymbol{Z}) - \lambda_d\left(\sum_{k=1}^{K}\eta_{dk}-1\right)\right] = \left[\sum_{i=1}^{n_d} Z_{dik}\right]\frac{1}{\eta_{dk}} - \lambda_d \quad (3)$$

$$\widehat{\eta}_{dk} = \frac{\sum_{i=1}^{n_d} Z_{dik}}{\lambda_d} \quad (4)$$

Sous la contrainte $\sum_{k=1}^{K}\eta_{dk}=1$, on obtient

$$\lambda_d = \sum_{k=1}^{K}\sum_{i=1}^{n_d} Z_{dik} = \sum_{i=1}^{n_d}\underbrace{\sum_{k=1}^{K} Z_{dik}}_{=1} = n_d$$

$$\widehat{\eta}_{dk} = \frac{\sum_{i=1}^{n_d} Z_{dik}}{n_d}$$

C'est la proportion du nombre de mots associés au topic k dans le texte d.

3. Rappeler le principe de l'algorihme EM et montrer qu'il génère une suite $\theta^{(h)}$ faisant croitre la vraisemblance [Question de cours]

4. Expliciter pour LDA la loi des variables latentes conditionnellement aux observations. On montrera au passage que les variables latentes $(Z_{di})_{i=,1,\ldots,n_d,d=1,\ldots,D}$ sont indépendantes conditionnellement aux observations. On notera $\tau_{dik} = P(Z_{di}=k|Y_{id})$.

$$\begin{aligned} p(\boldsymbol{Z}|\boldsymbol{Y}) &= \prod_{d=1}^{D}\prod_{i=1}^{n_d} \frac{p(Y_{di}|Z_{di})p(Z_{di})}{p(Y_{di})} \\ &= \prod_{d=1}^{D}\prod_{i=1}^{n_d} p(Z_{di}|Y_{di}) \end{aligned}$$

par indépendence des textes et des mots.

$$\begin{aligned} \tau_{dik} = P(Z_{di}=k|Y_{di}=y_{di}) &\propto P(Y_{di}=y_{di}|Z_{di}=k)P(Z_{di}=k) \\ &\propto \eta_{dk}\phi_{k y_{di}} \\ &\propto \eta_{dk}\sum_{v=1}^{V} y_{div}\phi_{kv} \\ C &= \sum_{k=1}^{K}\eta_{dk}\sum_{v=1}^{V} y_{div}\phi_{kv} \end{aligned}$$

5. Calculer $\mathbb{E}_{\theta^{(h)}}\left[\log p(\boldsymbol{Y},\boldsymbol{Z};\theta)|\boldsymbol{Y}=\mathbf{y}\right]$.

$$\mathbb{E}_{\theta^{(h)}}\left[\log p(\boldsymbol{Y},\boldsymbol{Z};\theta)|\boldsymbol{Y}=\mathbf{y}\right] = \sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{d=1}^{D}\sum_{i=1}^{n_d}\tau_{dik}^{(h)} y_{div}\right]\log\phi_{kv} + \sum_{d=1}^{D}\sum_{k=1}^{K}\left[\sum_{i=1}^{n_d}\tau_{dik}^{(h)}\right]\log\eta_{dk}$$

6. Ecrire l'étape M de l'algorihme EM pour le LDA. On pourra d'aider des calculs menés à la question 2.

$$\widehat{\eta}_{dk} = \frac{\sum_{i=1}^{n_d}\tau_{dik}}{n_d} \quad \text{and} \quad \widehat{\phi}_{kv} = \frac{\sum_{d=1}^{D}\sum_{i=1}^{n_d}\tau_{dik}Y_{div}}{\sum_{d=1}^{D}\sum_{i=1}^{n_d}\tau_{dik}}$$

7. Quel critère pouvez-vous proposer pour choisir le nombre de topics K?
8. Proposer une solution permettant de clusteriser les D textes en fonction des topics qu'ils abordent. Vous pouvez proposer un nouveau modèle ou bien juste un post-traitement des résultats obtenus après l'inférence.
9. Proposer une situation où ce modèle pourrait être utilisé en écologie, biologie ou sciences de l'environnement.

Part 2. Inférence Bayésienne (25%)

On propose de considérer une inférence bayésienne du modèle. On définit pour cela une loi a priori de Dirichlet sur les paramètres inconnus.

$$\begin{aligned}\boldsymbol{\eta}_d = (\eta_{d1}, \dots, \eta_{dK}) &\sim_{i.i.d} \mathcal{D}\text{ir}_K(\alpha, \dots, \alpha) \quad \forall d \in \{1, \dots, D\} \\ \boldsymbol{\phi}_k = (\phi_{k1}, \dots, \phi_{kV}) &\sim_{i.i.d} \mathcal{D}\text{ir}_V(\beta, \dots, \beta) \quad \forall k \in \{1, \dots, K\}\end{aligned}$$

On rappelle que la loi de Dirichlet de paramètres $(\alpha_1, \dots, \alpha_D)$ est une loi multivariée ayant pour support le simplexe $\mathcal{S}_D = \{(p_1, \dots, p_D) \in [0,1]^D \mid \sum_{d=1}^D p_d = 1\}$ de densité

$$f(p_1, \dots, p_D; \alpha_1, \dots, \alpha_D) = \text{Cste} \mathbb{1}_{\mathcal{S}_D}(p_1, \dots, p_D) \prod_{d=1}^D p_d^{\alpha_d - 1}$$

On cherche à générer un échantillon de $p(\theta, \boldsymbol{Z}|\boldsymbol{Y})$. Soit h le numéro de l'itération. On propose l'algorithme suivant.

A l'itération (h):
- a. Générer $\boldsymbol{\phi}_k^{(h)} \sim p(\boldsymbol{\phi}|\boldsymbol{Z}^{(h-1)}, \boldsymbol{Y}, \boldsymbol{\eta}^{(h-1)})$, $\quad \forall k = 1, \dots K$
- b. Générer $\boldsymbol{\eta}_d^{(h)} \sim p(\boldsymbol{\eta}|\boldsymbol{Z}^{(h-1)}, \boldsymbol{Y}, \boldsymbol{\phi}_k^{(h)})$, $\quad \forall d = 1, \dots D$
- c. Générer $\boldsymbol{Z}^{(h)} \sim p(Z|\boldsymbol{Y}, \boldsymbol{\phi}^{(h)}, \boldsymbol{\eta}^{(h)})$

10. [Question de cours] What is the name of the algorithm? Give quickly its properties.

11. Montrer que les étapes [a.] et [b.] reviennent à simuler des lois de Dirichlet dont on spécifiera les paramètres.

$$
\begin{aligned}
p(\boldsymbol{\phi}|\boldsymbol{Z},\boldsymbol{Y};\boldsymbol{\eta}) &\propto \exp\left[\sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}\right]\log\phi_{kv} + \sum_{d=1}^{D}\sum_{k=1}^{K}\left[\sum_{i=1}^{n_d} Z_{dik}\right]\log\eta_{dk}\right] p(\boldsymbol{\phi}) \\
&\propto \prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{kv}^{S_{kv}}\prod_{k=1}^{K}\prod_{v=1}^{V}\phi_{kv}^{\beta-1}\mathbb{1}_{\mathcal{S}_V}(\phi_{kv}) \\
&\propto \prod_{k=1}^{K}\mathbb{1}_{\mathcal{S}_V}(\phi_{kv})\prod_{v=1}^{V}\phi_{kv}^{S_{kv}+\beta-1} \\
\text{where } S_{kv} &= \sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}
\end{aligned}
$$

On reconnait un Dirichlet $\mathcal{D}(\beta + S_{k1}, \ldots, \beta + S_{kV})$.
Idem pour $\boldsymbol{\eta}$

$$
\begin{aligned}
p(\boldsymbol{\eta}|\boldsymbol{Z},\boldsymbol{Y};\boldsymbol{\phi}) &\propto \exp\left[\sum_{k=1}^{K}\sum_{v=1}^{V}\left[\sum_{d=1}^{D}\sum_{i=1}^{n_d} Z_{dik}Y_{div}\right]\log\phi_{kv} + \sum_{d=1}^{D}\sum_{k=1}^{K}\left[\sum_{i=1}^{n_d} Z_{dik}\right]\log\eta_{dk}\right] p(\boldsymbol{\eta}) \\
&\propto \prod_{d=1}^{D}\prod_{k=1}^{K}\eta_{dk}^{N_{dk}}\prod_{d=1}^{D}\prod_{k=1}^{K}\eta_{dk}^{\alpha-1}\mathbb{1}_{\mathcal{S}_K}(\boldsymbol{\eta}_d) \\
&\propto \prod_{d=1}^{D}\mathbb{1}_{\mathcal{S}_K}(\boldsymbol{\eta}_d)\prod_{k=1}^{K}\eta_{dk}^{N_{dk}+\alpha-1} \\
\text{where } N_{dk} &= \sum_{i=1}^{n_d} Z_{dik}
\end{aligned}
$$

Part 3. Pour aller plus loin (25%)

On s'intéresse aux emails envoyés par des individus. Si on indice par d les individus, le texte contenant l'ensemble des emails entre d et d' sera noté $\boldsymbol{Y}_{dd'}$ et le i-ème mot de ce texte sera $Y_{dd'i}$ pour $i = 1, \ldots, n_{dd'}$.

12. En combinant le LDA et le/les modèle(s) vu(s) en cours, proposez un modèle permettant de clusteriser les individus en fonction des topics qu'ils abordent dans leurs emails.

    □

13. Commentez la difficulté de l'inférence de ce modèle et proposez une solution envisageable (sans entrer dans les détails).