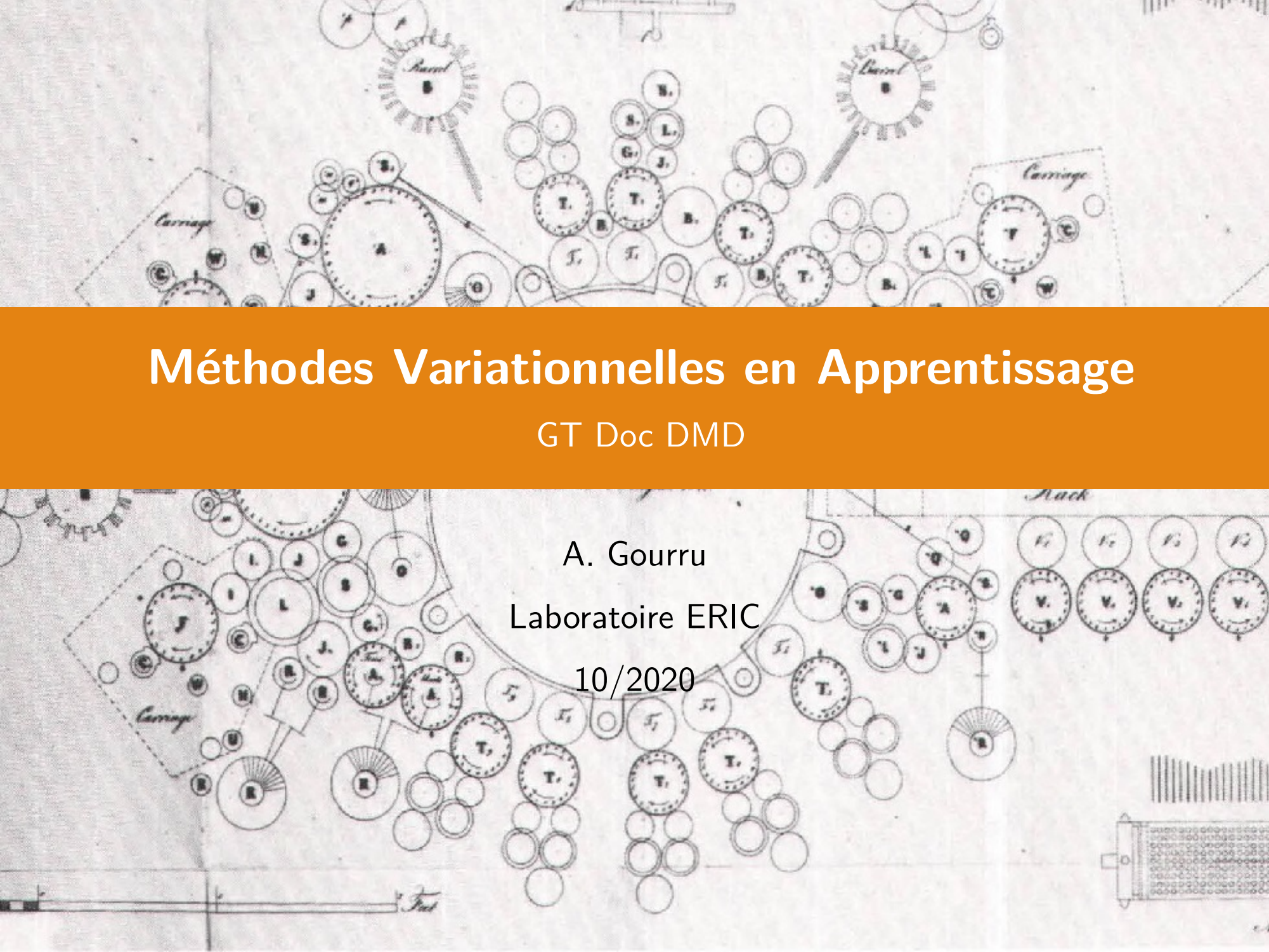


Méthodes Variationnelles en Apprentissage

GT Doc DMD

A. Gourru

Laboratoire ERIC

10/2020

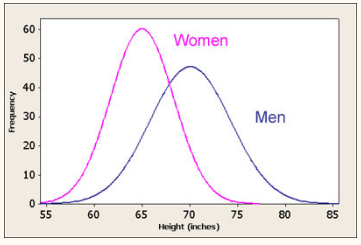
1 Sommaire

1 Introduction

- Cadre : modélisation probabiliste
- Les données sont issues d'un processus stochastiques + ou - complexe

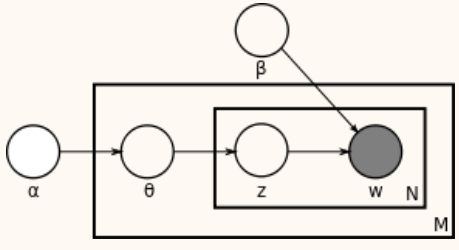
Exemple

La taille d'une population suit une loi normale (enfin...)



Exemple

Latent Dirichlet Allocation



On a deux ensembles :

- Les variables latentes, notées $z = z_{1:m}$, qui sont inconnues
- Les observations : $x = x_{1:n}$ issues de processus stochastiques paramétrés par les variables latentes

Exemple

1. La taille moyenne des fille μ_f et la taille moyenne des garçons μ_g sont des variables latentes, issues de deux tirages d'une même loi normale de paramètres μ_p et σ_p.
2. Pour chaque individu on tire son sexe α selon une Bernoulli de paramètre β, puis sa taille selon une normale avec comme paramètre la moyenne correspondant à son sexe et une variance globale σ_q latentes

Les données sont un ensembles de tailles. On cherche à identifier toutes les variables latentes $\{\mu_f, \mu_g, \mu_p, \sigma_p, \alpha, \beta, \sigma_q\}$

- Inférence Variationnelle [1]
  Déterminer la distribution $p(z|x)$, ou z est l'ensemble des variables latentes (v.l.). Problème :

$$p(z|x) = \frac{p(x,z)}{\int p(x,z)d_z}.$$

  Le dénominateur n'est généralement pas calculable. On va donc essayer d'apprendre une bonne approximation de $p(z|x)$

- Variational Information Bottleneck [2]
  On a en plus un ensemble de labels $y = y_{1:n}$ associés aux observations. Dans cette approche, les objectifs sont i) de compresser l'information dans les variables latentes, et ii) d'identifier les v.l. qui expliquent les labels :

$$\arg\max_z I(z,y) - \beta I(z,x) \tag{1}$$

  Ou I est l'information mutuelle, qui nécessite le calcule de $p(z|x)$ notamment.

2 Sommaire

On cherche à déterminer la distribution $p(z|x)$, ou z est un ensemble de variables latentes.

$$p(z|x) = \frac{p(x,z)}{\int p(x,z)d_z}.$$

Le dénominateur est intraitable dans la majorité des cas [1]
Le but de l'inférence variationnelle est d'approximer cette distribution par une autre distribution $q_\lambda(z)$. En choisissant Kullback liebler comme dissimilarité entre les distributions, on veut minimiser en lambda :

$$KL(q_\lambda(z)||p(z|x)) = \int q_\lambda(z) \log \frac{q_\lambda(z)}{p(z|x)} d_z \tag{2}$$

En rappelant que $p(z|x) = \frac{p(x,z)}{p(x)}$, on a :

$$
\begin{aligned}
KL(q_\lambda(z)||p(z|x)) &= \int q_\lambda(z) \log \frac{q_\lambda(z)}{p(z|x)} d_z \\
&= \int q_\lambda(z) \log \frac{q_\lambda(z)p(x)}{p(x,z)} d_z \\
&= \int q_\lambda(z) \log \frac{q_\lambda(z)}{p(x,z)} d_z + \int q_\lambda(z) \log p(x) d_z \\
&= \int q_\lambda(z) \log \frac{q_\lambda(z)}{p(x,z)} d_z + \log p(x) \int q_\lambda(z) d_z \\
&= \int q_\lambda(z) \log \frac{q_\lambda(z)}{p(x,z)} d_z + \log p(x) \\
&= -\mathcal{L}(q) + \log p(x)
\end{aligned}
\tag{3}
$$

$\mathcal{L}$ est appelé la ELBO. On obtient une décomposition de $\log p(x)$

$$\log p(x) = \mathcal{L}(q) + KL(q_\lambda(z)||p(z|x)) \tag{4}$$

$\log p(x)$ étant fixe, et $KL(q_\lambda(z)||p(z|x))$ positive, maximiser $\mathcal{L}(q)$ revient à minimiser $KL(q_\lambda(z)||p(z|x))$, et donc rapprocher $q_\lambda(z)$ de $p(z|x)$.

$$\begin{aligned}
\mathcal{L}(q) &= -\int q_\lambda(z) \log \frac{q_\lambda(z)}{p(x,z)} d_z \\
&= \int q_\lambda(z) \log p(x,z) d_z - \int q_\lambda(z) \log q_\lambda(z) \\
&= \mathbb{E}_q[\log p(x,z)] - \mathbb{E}_q[\log q_\lambda(z)]
\end{aligned} \tag{5}$$

Lorsque l'espérance n'est pas calculable en forme close (famille non exponentielle, loi non conjuguées), on peut avoir recours à une approximation de cette espérance par une méthode de Monte Carlo. Plusieurs articles traitent de cette approche tout en proposant des façons de réduire la variance de l'estimateur [3, 4].
L'approche consiste à réaliser L tirages aléatoires de z suivant q

$$\mathbb{E}_q[\log p(x|z)] \approx \frac{1}{L}\sum_{l=1}^{L} \log p(x|z^{(l)}),\ z^{(l)} \sim q_\lambda(z) \tag{6}$$

On peut agir similairement pour la deuxième partie de la ELBO. Elle est néanmoins bien souvent calculable en forme close : en considérant sa valeur exacte, on réduit la variance et le temps de calcul.

Cette méthode proposée dans [4] réduit empiriquement la variance de l'estimateur dans de nombreux cas [**?**] (pas de preuve théorique à ma connaissance) et rend possible l'optimization par Réseau de Neurone
Dans certains cas, la densité $q_\lambda(z)$ peut être "reparamétrée" au moyen d'une fonction $g_\lambda(x, \epsilon)$ différentiable sur x et d'une variable aléatoire auxiliaire ϵ [4]. On va remplacer

$$z \sim q_\lambda(z) \tag{7}$$

par

$$z = g_\lambda(x, \epsilon), \quad \epsilon \sim p(\epsilon) \tag{8}$$

Dans le cas d'un loi normale avec variance diagonale par exemple, on a

$$z \sim q_\lambda(z) \iff z = \mu_\lambda(x) + \sigma^2_\lambda(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, 1) \tag{9}$$

Ici, μ_λ et σ^2_λ sont des fonctions de x paramétrées par les paramètres variationnels λ (voir section **??**).

Dans le cas ou l'espérance de la vraisemblance est approximée, la ELBO pour être vue comme une fonction objective comprenant une erreur de reconstruction et une régularisation forçant la densité variationnelle à être proche de la densité à priori des paramètres latents.

$$\mathcal{L}(q) = \frac{1}{L} \sum_{l=1}^{L} \log p(x|z^{(l)}) - KL(q_\lambda(z)||p(z)) \tag{10}$$
$$z^{(l)} \sim q_\lambda(z)$$

Pourquoi une erreur de reconstruction ? On génère un "code" à partir d'une observation, et on essaye de maximiser la probabilité de l'observation reconstruite à partir de ce code. Le parallèle avec un autoencoder est entériné par [4].

3 Sommaire

3 VIB

Proposé pour la première fois par [5]. Le principe de cette méthode est d'apprendre une représentation z qui maximise la compression des données initiales x, tout en étant informative par rapport au labels y associés aux données.

$$\arg\max_z I(z,y) - \beta I(z,x) \tag{11}$$

ou I est l'information mutuelle.

$$I(x,y) = \int\int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} d_x d_y \tag{12}$$

Le paramètre $\beta \geq 0$ contrôle l'équilibre entre ces deux sous objectifs. Une valeur β élevée entraîne une représentation hautement compressée.
Le premier terme de d'équation 11 encourage z à bien prédire y ; le deuxième terme encourage z à «oublier» x de façon à le compresser. Essentiellement, cela force z à agir comme une statistique minimale suffisante de x pour prédire y.

L'information mutuelle n'est généralement pas calculable en forme close. [2] propose d'utiliser une approche variationnelle [1]. Associée à l'astuce de reparamétrisation [4] permettant d'estimer efficacement l'équation 11, l'équation est maximisée en utilisant une architecture particulière de réseau de neurones.
Ici le principe est un peu différent : p(z|x) est un choix de modélisation, et on va approximer p(z), et p(y|z).
On construit une borne inférieure qui est :

$$-L_{VIB} = \mathbb{E}_{z\sim p(z|x)}[\log q(y|z)] - \beta \mathbf{KL}(p(z|x)||r(z)) \tag{13}$$

ou $p(z|x)$ est l'encoder, $q(y|z)$ est le decoder, et r(z) est approximé par une gaussienne centré réduite $N(0, I)$. L'espérance est approximée par une méthode de tirage, similairement à [4].

- Travail publié à ICLR 2019, [6] pour les images.

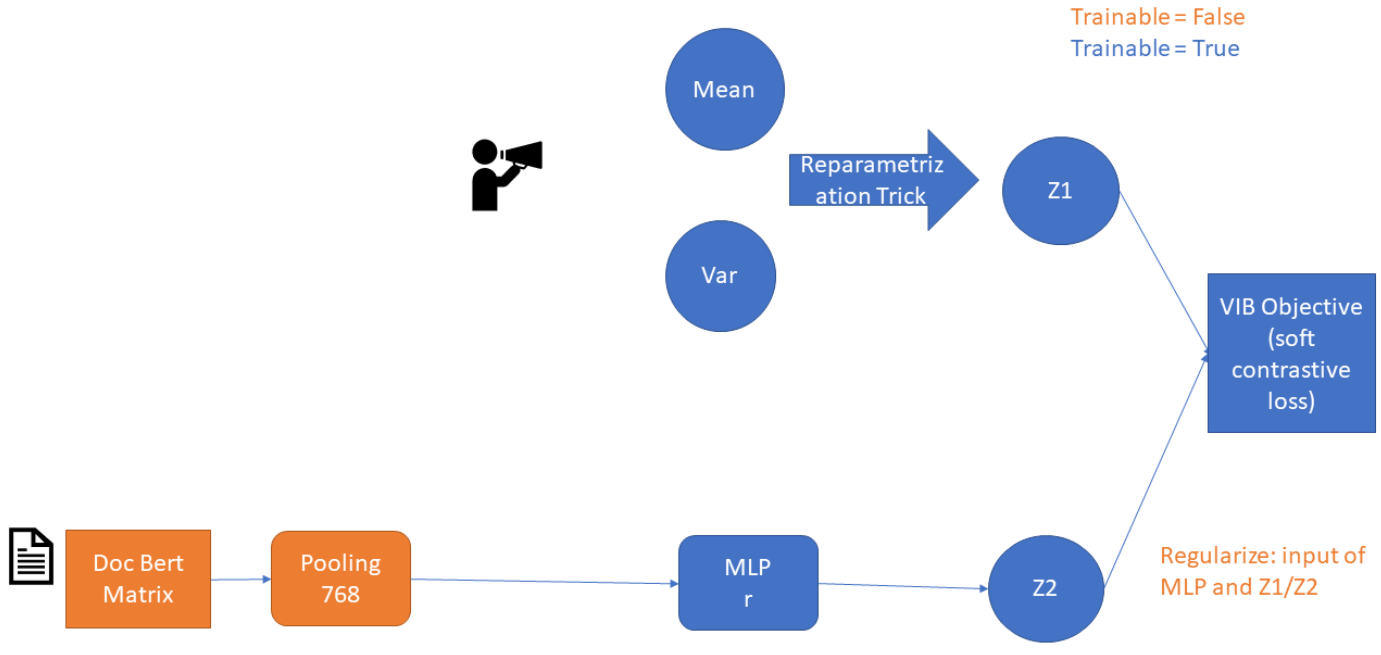
$y = 1$ si deux images sont de la mm classe (ex : MNIST). On tire un ensemble d'exemples négatif (de paires d'images qui ne sont pas dans la mm classe), pour qui $y = 0$.
La probabilité du label pour une pair d'image (a,b) est

$$q\left(y|z_a, z_b\right) = \sigma\left(-c||z_a - z_b||_2 + e\right) \tag{14}$$

avec σ la fonction sigmoïde et $c > 0$ et $c \in \mathbb{R}$ sont des paramètres scalaires. $p(z_a|x_a)$ est un réseau de neurones (VAE).

$$\begin{aligned} L_{VIBEmb} = -\mathbb{E}_{p(z_a|x_a),p(z_b|x_b)})[\log q\left(y = \hat{y}|z_a, z_b\right)] \\ -\beta[\mathbf{KL}(p(z_a|x_a)||r(z_a)) + \mathbf{KL}(p(z_b|x_b)||r(z_b))] \end{aligned} \tag{15}$$

$y = 1$ si vraie paire auteur-document $y = 0$ sinon.



Merci

3 Références I

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul.
An introduction to variational methods for graphical models.
Machine learning, 37(2):183–233, 1999.

Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy.
Deep variational information bottleneck.
arXiv preprint arXiv:1612.00410, 2016.

Rajesh Ranganath, Sean Gerrish, and David Blei.
Black box variational inference.
In *Artificial Intelligence and Statistics*, pages 814–822, 2014.

Diederik P Kingma and Max Welling.
Auto-encoding variational bayes.
arXiv preprint arXiv:1312.6114, 2013.

3 Références II

Naftali Tishby, Fernando C Pereira, and William Bialek.
The information bottleneck method.
The 37th annual Allerton Conference on Communication, Control, and Computing, page 368–377, 1999.

Seong Joon Oh, Kevin Murphy, Jiyan Pan, Joseph Roth, Florian Schroff, and Andrew Gallagher.
Modeling uncertainty with hedged instance embedding.
arXiv preprint arXiv:1810.00319, 2018.