Introduction aux modèles graphiques **2010/2011**

Cours 1 — 1 octobre

Enseignant: Francis Bach *Scribe: Guillaume Tartavel, Charles Miglietti*

<u>Pour information</u>

– Page web du cours `http://www.di.ens.fr/~fbach/courses/fall2010/`

1.1 Rappels de proba

1.1.1 Notations

Soit $\{X_1, \cdots, X_n\}$ un ensemble de variables aléatoires, de distribution :

$$\mathbb{P}\left(X_1 = x_1, \cdots, X_p = x_p\right) = p\left(x_1, \cdots, x_p\right)$$

.

On note (de manière ambiguë) $\mathbb{P}\left(X_i = x_i\right) = p\left(x_i\right)$.
Pour $A, B \in \{1, \cdots, p\}$, on note $X_A = \left(X_i\right)_{i \in A}$ et $X_B = \left(X_j\right)_{j \in B}$.

1.1.2 Quelques définitions / formules

- Loi marginale : $p\left(x_A\right) = \sum_{x_{A^c}} p\left(x_A, x_{A^c}\right)$
- Loi conditionnelle : $p\left(x_A | x_B\right) = \frac{p(x_A, x_B)}{p(x_B)}$ si $p\left(x_B\right) \neq 0$
- Formule de Bayes : $p\left(x_A | x_B\right) = \frac{p(x_B | x_A) p(x_A)}{p(x_B)}$

1.1.3 Espérances

- Espérance de X_i : $\mathbb{E}\left[X_i\right] = \sum_{x_i} x_i \cdot p\left(x_i\right)$
- Espérance de $f\left(X_i\right)$, pour f mesurable : $\mathbb{E}\left[f\left(X_i\right)\right] = \sum_{x_i} f\left(x_i\right) \cdot p\left(x_i\right)$
- Variance :

$$\begin{aligned} Var\left(X_i\right) &= \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)^2\right] \\ &= \mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[X_i\right]^2 \end{aligned}$$

- Covariance :

$$\begin{aligned} Cov\left(X_i, X_j\right) &= \mathbb{E}\left[\left(X_i - \mathbb{E}\left[X_i\right]\right)\left(X_j - \mathbb{E}\left[X_j\right]\right)\right] \\ &= \mathbb{E}\left[X_i X_j\right] - \mathbb{E}\left[X_i\right]\mathbb{E}\left[X_j\right] \end{aligned}$$

1.1.4 Indépendance

- Indépendance de 2 variables :
  X_i, X_j indépendantes $\Leftrightarrow p(x_i, x_j) = p(x_i)\, p(x_j) \quad \forall x_i, x_j$.
  On note $X_i \perp X_j$.
- Indépendance de n variables :
  $X_1, \cdots, X_n$ indépendantes $\Leftrightarrow p(x_1, \cdots, x_n) = p(x_1) \cdots p(x_n) \quad \forall x_1, \cdots, x_n$.
- Attention : indépendance de n variables $\Rightarrow$ indépendance 2 à 2,
  mais *la réciproque est fausse* !

1.1.5 Espérances et indépendance conditionnelles

En considérant la distribution (conditionnelle) $p_{x_j}(x_i) = p(x_i|x_j) = \frac{p(x_i,x_j)}{p(x_j)}$ au lieu de $p(x_i)$, on obtient les notions d'espérances conditionnelles.

- Espérance conditionnelle de X_i : $\mathbb{E}[X_i|X_j] = \sum_{x_i} x_i \cdot p(x_i|x_j)$
- Variance conditionnelle :

$$\begin{aligned} Var(X_i) &= \mathbb{E}\left[(X_i - \mathbb{E}[X_i|X_j])^2 |X_j\right] \\ &= \mathbb{E}\left[X_i^2|X_j\right] - \mathbb{E}[X_i|X_j]^2 \end{aligned}$$

- Indépendance conditionnelle :
  X_i, X_k indépendantes conditionnellement à $X_j \Leftrightarrow p(x_i, x_k|x_j) = p(x_i|x_j)\, p(x_k|x_j) \quad \forall x_i, x_j, x_k$
  On note $X_i \perp X_k | X_j$.

Exemples et contres-exemples : on réalise deux lancers d'une pièce de monnaie. Soient les variables aléatoire X_1, X_2 valant 1 si le lancer donne pile, 0 sinon, et $X_3 = XOR(X_1, X_2)$. On montre alors que :

- les $(X_i)_{i=1,2,3}$ sont indépendantes 2 à 2 ;
- les $(X_i)_{i=1,2,3}$ ne sont pas indépendantes ;
- deux des $(X_i)_{i=1,2,3}$ ne sont pas indépendantes sachant la troisième.

1.2 Modèles à un noeud

1.2.1 Modèle ?

Notations

$X_1, \cdots, X_n$ désignent des variables aléatoires IID (Indépendantes et Identiquement Distribuées) ;
$x_1, \cdots, x_n$ sont des observations de X ;
$\{x_1, \cdots, x_n\}$ est l'échantillon observé.

Définition

Un modèle est un ensemble de distributions avec des paramètres : $\{p_\theta(x), \theta \in \Theta\}$, avec généralement $\Theta = \mathbb{R}^p$.

Par exemple : pour une gaussienne à une dimension, les paramètres sont la moyenne et la variance : $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$.

L'objectif est de déterminer (d'inférer) les paramètres θ du modèle à partir d'observations (de réalisations) de ce dernier.

Approche fréquentiste

Les paramètres optimaux θ^* sont définis comme extrêmum d'une fonction de contraste (ou fonction de coût / de perte).

$\hat{\theta} = T(X_1, \cdots, X_n)$ étant un estimateur, le but est de faire tendre $\hat{\theta}$ vers θ^*.

Approche bayésienne

On fait un *a priori* $p(\theta)$ sur les paramètres que peut prendre le modèle.

Les observations sont des probabilités, connaissant les paramètres : $p(x|\theta)$.

La probabilité *a posteriori* du modèle est alors $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \propto p(x|\theta)\, p(\theta)$.

Maximum de vraisemblance (Fisher)

La vraisemblance (*likelihood*) d'un modèle est la quantité $L(\theta) = p_\theta(x_1, \cdots, x_n)$, vue comme fonction de θ.

On utilise souvent la log-vraisemblance, $l(\theta) = \log L(\theta)$.

L'estimateur du maximum de vraisemblance est le maximiseur de la vraisemblance :

$$\hat{\theta} = \arg\max_{\theta \in \Theta} L(\theta)$$

1.2.2 Modèles élémentaires

Loi de Bernouilli

Définition

$X \in \{0, 1\}$ est une variable aléatoire et $\theta \in [0, 1]$ un paramètre.

X suit une loi de Bernouilli, $X \sim \mathrm{Ber}(\theta)$, si :

$$\begin{cases} p(X = 1) & = \theta \\ p(X = 0) & = 1 - \theta \end{cases}$$

Une autre manière de l'écrire :

$$p(x) = \theta^x (1 - \theta)^{(1-x)} \tag{1.1}$$

Propriétés

- $\mathbb{E}[X] = \theta$
- $\mathrm{Var}(X) = \theta(1-\theta)$
- Si $X_1, \cdots, X_n \sim \mathrm{Ber}(\theta)$ iid, alors $Z = \sum_i X_i \sim \mathrm{Binom}(n, p)$

Maximum de vraisemblance

Soient $X_1, \cdots, X_n$ iid de loi $\mathrm{Ber}(\theta)$.
Vraisemblance :

$$L(\theta) = p_\theta(x_1, \cdots, x_n) = \prod_{i=1}^{n} p_\theta(x_i)$$

Log-vraisemblance, en utilisant l'expression (**??**) :

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log p_\theta(x_i) \\ &= \sum_{i=1}^{n} \left(x_i \log\theta + (1-x_i)\log(1-\theta)\right) \\ &= \left(\sum_{i=1}^{n} x_i\right)\log\theta + \left(n - \sum_{i=1}^{n} x_i\right)\log(1-\theta) \end{aligned}$$

Cette expression est dérivable :

$$\frac{\partial l}{\partial\theta}(\theta) = \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} x_i\right)$$

Cette log-vraisemblance étant concave, son maximum est atteint quand la dérivée s'annule, d'où :

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Loi multinomiale

Définition

$X \in \{1, \cdots, q\}$ est une variable aléatoire.
Les paramètres sont $\Pi = (\pi_1, \cdots, \pi_q) \in \mathbb{R}_+^q$, qui vérifient :

$$\sum_k \pi_k = 1 \tag{1.2}$$

La distribution d'une loi multinomiale $X \sim \mathrm{Multi}(1, \pi_1, \cdots, \pi_q)$ est : $p_\Pi(k) = \pi_k \quad \forall k$.

Une autre manière de l'écrire :

$$p_\Pi(x) = \prod_{k=1}^{q} \pi_k^{\delta(x,k)} \tag{1.3}$$

Pour $X_i \sim \text{Multi}(1, \pi_1, \cdots, \pi_q)$, on définit $\Delta_i = (\delta_{i,1}, \cdots, \delta_{i,q})$ avec $\delta_{i,k} = \delta(X_i, k) \quad \forall k$.

Propriétés

Si $X_1, \cdots, X_n \sim \text{Multi}(1, \pi_1, \cdots, \pi_q)$ iid, alors :

$$Z' = \sum_{i=1}^{n} \Delta_i \sim \text{Multi}(n, \pi_i, \cdots, \pi_q)$$

Et :

$$p(Z' = (n_1, \cdots, n_q)) = \frac{n!}{n_1! \cdots n_q!} \prod_{k=1}^{q} \pi_k^{n_k}$$

Maximum de vraisemblance

Soient $X_1, \cdots, X_n$ iid de loi $Multi(1, \pi_1, \cdots, \pi_k)$ (et donc $\Delta_1, \cdots, \Delta_n$ sont iid).
Vraisemblance :

$$L(\Pi) = p_\Pi(x_1, \cdots, x_n) = \prod_{i=1}^{n} p_\Pi(x_i)$$

Log-vraisemblance, utilisant l'expression (**??**) :

$$\begin{aligned} l(\Pi) &= \sum_{i=1}^{n} \log p_\Pi(x_i) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{q} (\delta_{i,k} \log \pi_k) \\ &= \sum_{k=1}^{q} n_k \log \pi_k \end{aligned}$$

Où on a posé $n_k = \sum_i \delta_{i,k}$ le nombre d'observations ayant pour valeur k.
Pour prendre en compte la contrainte (**??**), considérons le lagrangien :

$$\mathcal{L}(\Pi, \lambda) = -\underbrace{\sum_{k=1}^{q} n_k \log \pi_k} + \lambda \left(\sum_{k=1}^{q} \pi_k - 1 \right)$$

L'objectif est de maximiser la log-vraisemblance sous la contrainte (**??**), ce qui revient à chercher :

$$\min_{\Pi \in \mathbb{R}^q} \left[\max_{\lambda \in \mathbb{R}_+} \mathcal{L}(\Pi, \lambda) \right]$$

$\mathcal{L}$ étant convexe pour Π, ce problème est équivalent à :

$$\max_{\lambda \in \mathbb{R}_+} \left[\min_{\Pi \in \mathbb{R}^q} \mathcal{L}(\Pi, \lambda) \right]$$

Dérivons alors $\mathcal{L}$ par rapport aux composantes de Π :

$$\frac{\partial \mathcal{L}}{\partial \pi_k}(\Pi, \lambda) = -\frac{n_k}{\pi_k} + \lambda$$

Le minimum de $\mathcal{L}$ par rapport à Π est atteint en annulant cette expression $\forall k$, d'où : $n_k = \lambda \pi_k$.

En sommant cette relation pour $k = 1 \cdots q$ et en utilisant la contrainte (**??**), il vient que $n = \lambda$, et la relation précédente donne donc donc :

$$\hat{\pi_k} = \frac{n_k}{n}$$

Loi gaussienne (1D)

Définition

X est une variable aléatoire réelle, et $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+^*$ sont les paramètres.
X suit une loi gaussienne, $X \sim \mathcal{N}(\mu, \sigma^2)$, signifie :

$$p_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right)$$

Maximum de vraisemblance

Soient $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ iid.
Vraisemblance :

$$L\left(\mu, \sigma^2\right) = p_{\mu,\sigma^2}(x_1, \cdots, x_n) = \prod_{i=1}^n p_{\mu,\sigma^2}(x_i)$$

Log-vraisemblance :

$$\begin{aligned} l\left(\mu, \sigma^2\right) &= \sum_{i=1}^n \log p_{\mu,\sigma^2}(x_i) \\ &= -\frac{n}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \end{aligned}$$

Dérivées, à annuler pour obtenir le maximum :

$$\begin{aligned}
\frac{\partial l}{\partial \mu}\left(\mu, \sigma^2\right) &= \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right) \\
&= \frac{1}{\sigma^2}\left(\sum_{i=1}^{n} x_i - n\mu\right) \\
\frac{\partial l}{\partial \sigma^2}\left(\mu, \sigma^2\right) &= -\frac{n}{2}\frac{2\pi}{2\pi\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \\
&= -\frac{1}{2\sigma^4}\left(n\sigma^2 - \sum_{i=1}^{n}\left(x_i - \mu\right)^2\right)
\end{aligned}$$

Ce qui nous donne :

$$\begin{aligned}
\hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
\hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)^2
\end{aligned}$$

Loi gaussienne (kD)

Définition

X est une variable aléatoire à valeur dans $\mathbb{R}^d$.
$\theta = (\mu, \Sigma) \in \mathbb{R}^d \times \mathbb{R}^{d\times d}$ sont les paramètres (avec Σ définie positive).
X suit une loi gaussienne, $X \sim \mathcal{N}(\mu, \Sigma)$, signifie :

$$p_{\mu,\Sigma}(x) = \frac{1}{\sqrt{(2\pi)^d \det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

Maximum de vraisemblance

Soient $X_1, \cdots, X_n \sim \mathcal{N}(\mu, \Sigma)$ iid.
Log-vraisemblance :

$$\begin{aligned}
-l(\mu, \Sigma) &= -\sum_{i=1}^{n} \log p_{\mu,\Sigma}(x_i) \\
&= \frac{nd}{2}\log(2\pi) - \frac{n}{2}\log\left(\det \Sigma^{-1}\right) + \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)^T \Sigma^{-1}\left(x_i - \mu\right)
\end{aligned}$$

Par convexité, le minimum de $-l(\mu, \Sigma)$ est le point en lequel son gradient s'annule.

Gradient par rapport à μ :

$$\begin{aligned} -\nabla_{\mu} l\left(\mu, \Sigma\right) &= \sum_{i=1}^{n} \Sigma^{-1}\left(x_i - \mu\right) \\ &= \Sigma^{-1}\left(\sum_{i=1}^{n} x_i - n\mu\right) \end{aligned}$$

Gradient par rapport à Σ^{-1} :

$$-\nabla_{\Sigma^{-1}} l\left(\mu, \Sigma\right) = -\frac{n}{2}\left(\Sigma^{-1}\right)^{-1} + \frac{1}{2}\sum_{i=1}^{n}\left(x_i - \mu\right)\left(x_i - \mu\right)^{T}$$

D'où :

$$\begin{aligned} \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ \hat{\Sigma} &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu\right)\left(x_i - \mu\right)^{T} \end{aligned}$$