Entropie, KL, MV, Familles exponentielles **2010/2011**

Cours 5 — 27 octobre

Enseignant: Francis Bach *Scribe: Vincent Adam, Samy Blusseau*

5.1 Entropie

Définition 5.1 *Soit X une variable aléatoire dans $\mathcal{X}$ fini. On note $p(x) = P(X = x)$. L'entropie de X est définie par*

$$H(X) = -\sum_{x\in\mathcal{X}} p(x)\ \log\ p(x) = E_{p(x)}\frac{1}{\log\ p(x)}$$

Proposition 5.2 *On a les inégalités suivantes :*

1. $H(X) \geq 0$ *avec égalité si X est constant presque surement*
2. $H(X) \leq \log(Card(\mathcal{X}))$

$\hookrightarrow$ **Preuve :**

1. En prolongeant $p \to p\ \log\ p$ en 0 par 0, on a $\forall x \in \mathcal{X}, p(x)\ \log\ p(x) \geq 0$ d'ou $H(X) \geq 0$. Et si $\exists x_i \in \mathcal{X}$ tel que $p(x_i) = 1$ alors $H(X) = 0$
2. Par concavité de la fonction logarithme, l'inégalité de Jensen donne le résultat.

$$\begin{aligned} H(X) = E_{p(x)}\frac{1}{\log\ p(x)} &\leq \log\left(E_{p(x)}\frac{1}{p(x)}\right)\ (Jensen) \\ &\leq \log\left(\sum_{x\in\mathcal{X}}\frac{p(x)}{p(x)}\right) \\ &\leq \log(Card(\mathcal{X})) \end{aligned}$$

5.2 Divergence de Kullback-Leibler

Définition 5.3 *Divergence de Kullback Leibler*
Soient p et q deux distributions sur $\mathcal{X}$ finies. La divergence de Kullback Leibler entre p et q est définie par

$$D(p\ \|\ q) = \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} = E_{p(x)}\log\left(\frac{p(x)}{q(x)}\right)$$

La divergence n'est pas symétrique, ce n'est pas une distance

Proposition 5.4 *$D(p \parallel q) \geq 0$ avec égalité ssi $p = q$*

$\hookrightarrow$ **Preuve :** Par concavité de la fonction logarithme, l'inégalité de Jensen donne le résultat.

$$D(p \parallel q) = E_{p(x)}\left[-\log\left(\frac{q(x)}{p(x)}\right)\right] \geq -\log\left(E_{p(x)}q(x)\right) \geq 0$$

Définition 5.5 *Soit X, Y deux variables aléatoires de loi jointe $p(x, y) = P(X = x, Y = y)$, l'information mutuelle de X et Y est*

$$I(X, Y) = D(p(x, y) \parallel p(x)p(y))$$

Proposition 5.6 *$I(X, Y) \geq 0$ avec égalité ssi $X \perp Y$*

$\hookrightarrow$ **Preuve :** Par positivité de la divergence KL et définition de l'indépendance entre X et Y : $p(x, y) = p(x)p(y)$

Indépendance $\Rightarrow$ décorrélation **mais** décorrélation $\not\Rightarrow$ Indépendance
En effet, si $X \perp Y$ alors $E(X, Y) = E(X)E(Y)$ et donc $Cov(X, Y) = 0$.
Contre-exemple : si C le carré défini par $\mid x - y \mid \leq 1$ et p la densité uniforme sur ce carré $p((x, y) \in C) = 1/2$ On a $Cov(X, Y) = 0$, mais $p(x, y) \neq p(x)p(y)$

Remarque : La réciproque n'est vraie que dans le cas des variables aléatoires gaussiennes.

5.3 Lien entre la divergence de Kullback et le maximum de vraisemblance

Définition 5.7 *Soient $x_1, ..., x_N \in \mathcal{X}$ N observations i.i.d d'une variable aléatoire X. La loi empirique de X construite à partir de ces observations est*

$$\hat{p}(x) = \frac{1}{N}\sum_{n=1}^{N}\delta(x - x_n)$$

Où δ est la fonction dirac, nulle partout sauf en 0 où elle vaut 1 (cas discret)

Proposition 5.8 *Soit p_θ une distribution paramétrique sur $\mathcal{X}$.*
Maximiser la vraisemblance $p_\theta(x)$ revient à minimiser la divergence $D(\hat{p}||p_\theta)$

↪ **Preuve :**

$$\begin{aligned} D(\hat{p}||p_\theta) &= \sum_{x\in\mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} \\ &= H(\hat{p}) - \sum_{x\in\mathcal{X}} \hat{p}(x)\ \log\ p_\theta(x) \\ &= H(\hat{p}) - \sum_{x\in\mathcal{X}} \sum_{n=1}^{N} \delta(x - x_n)\ \log\ p_\theta(x) \\ &= H(\hat{p}) - \frac{1}{N} \sum_{n=1}^{N} \log\ p_\theta(x_i) \end{aligned}$$

Le second terme est égal à l'opposé de la vraisemblance $p_\theta(x)$. D'où la conclusion.

5.4 Lien entre entropie et divergence de Kullback-Leibler

Proposition 5.9 *Soit X une variable aléatoire sur $\mathcal{X}$ de distribution p et $unif$ la distribution uniforme sur $\mathcal{X}$, alors*

$$D(p\|unif) = -H(X) + \log(Card(\mathcal{X}))$$

.

↪ **Preuve :** $unif(x) = \frac{1}{Card(\mathcal{X})}$, d'où $D(p||unif) = \sum_{x\in\mathcal{X}} p(x)\ \log \frac{p(x)}{\frac{1}{Card(\mathcal{X})}}$

Remarque : Dans le cas discret, la distribution uniforme maximise l'entropie. La divergence donne ici l'écart entre l'entropie de p et l'entropie maximale réalisable.

Définition 5.10 *Entropie différentielle Soit $X \in \mathbb{R}^p$ une variable aléatoire de densité $p(x)$ par rapport à la mesure de Lebesgue. L'entropie différentielle de X est définie par*

$$H(X) = \int_{\mathbb{R}} p(x)\ \log\ p(x)\, dx$$

Remarque : L'entropie n'est pas invariante par changement de mesure.

5.5 Familles exponentielles

Définition 5.11 *Famille exponentielle : Soit X une variable aléatoire sur $\mathcal{X}$. Une famille exponentielle est définie par :*

- *Une mesure de référence $h(x)dx$*
- *Des descripteurs $\varphi(x) \in \mathbb{R}^p$, encore appelés "features" ou plus communément "sufficient statistics"*
- *Un paramètre naturel $\eta \in \mathbb{R}^p$*
- *Une fonction de log-partition $A(\eta)$*

tels que la densité de X s'écrit

$$p(x|\eta) = h(x) \exp\left\{\eta^T \varphi(x) - A(\eta)\right\}$$

Proposition 5.12

$$A(\eta) = \log \int_{\mathcal{X}} h(x) \exp\left\{\eta^T \varphi(x)\right\} dx$$

$\hookrightarrow$ **Preuve :**

$$1 = \int_{\mathcal{X}} p(x|\eta)dx \;=\; e^{-A(\eta)} \int_{\mathcal{X}} h(x) \exp\left\{\eta^T \varphi(x)\right\} dx$$

Définition 5.13 *On définit le Domaine par :*

$$Domaine \;= \{\eta \in \mathbb{R}^p, A(\eta) < \infty\}$$

Exemple 5.5.1 *Loi de Bernouilli : $\mathcal{X} = \{0, 1\}$, $p(x = 1) = \pi$*

$$\begin{aligned} p(x) &= \pi^x (1-\pi)^{1-x} \\ &= \left(\frac{\pi}{1-\pi}\right)^x (1-\pi) \\ &= \exp\left\{x \, \log \frac{\pi}{1-\pi}\right\} \exp\left\{\log(1-\pi)\right\} \end{aligned}$$

*On retrouve bien une famille exponentielle en posant $\eta = \log \frac{\pi}{1-\pi}$ (**log odd ratio**) et $A(\eta) = -\log(1-\pi) = \log(1+e^\eta)$:*

$$p(x) = e^{x\eta - A(\eta)}$$

Et le domaine est $\mathbb{R}$.

Exemple 5.5.2 *Loi Gaussienne(μ, σ) sur $\mathbb{R}$:*

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}|\sigma|} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\log \sigma^2 - \frac{x^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} + \frac{x\mu}{\sigma^2}\right\} \end{aligned}$$

On reconnait une famille exponentielle avec $\varphi(x) = (x,\, x^2)^T$, $\eta = (\frac{\mu}{\sigma^2},\, -\frac{1}{2\sigma^2})^T = (\eta_1,\, \eta_2)^T$ *et* $A(\eta) = \frac{1}{2}\log\,\sigma^2 + \frac{\mu^2}{2\sigma^2} = \frac{1}{2}\log\left(-\frac{1}{2\eta_2}\right) - \frac{\eta_1^2}{4\eta_2}$ *:*

$$p(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{\varphi(x)^T\eta - A(\eta)\right\}$$

avec pour domaine : $\{\eta \in \mathbb{R}^2, \eta_2 < 0\}$

Exemple 5.5.3 *Autres lois qui sont des familles exponentielles : Loi multinomiale, loi de Poisson (*$\mathcal{X} = \mathbb{N}$*), loi de Dirichlet, loi Gamma, loi exponentielle.*

5.5.1 Liens entre moments et $A(\eta)$

Il existe des liens entre les dérivées de $A(\eta)$ et les moments des familles exponentielles. En effet

$$\begin{aligned}
\nabla A(\eta) &= \frac{\int_{\mathcal{X}} h(x)\varphi(x)e^{\eta^T\varphi(x)}dx}{\int_{\mathcal{X}} h(x)e^{\eta^T\varphi(x)}dx} \\
&= \frac{\int_{\mathcal{X}} h(x)\varphi(x)e^{\eta^T\varphi(x)}dx}{e^{A(\eta)}} \\
&= \int_{\mathcal{X}} h(x)e^{\eta^T\varphi(x)-A(\eta)}\varphi(x)dx \\
&= \int_{\mathcal{X}} p(x|\eta)\varphi(x)dx \\
&= \mathbb{E}_{X|\eta}\,\varphi(X)
\end{aligned}$$

et (en utilisant le fait que $\nabla(e^{-A}) = -e^{-A}\nabla A$)

$$\begin{aligned}
\nabla^2 A(\eta) &= e^{-A(\eta)}\int_{\mathcal{X}} h(x)e^{\eta^T\varphi(x)}\varphi(x)\varphi(x)^T dx + \int_{\mathcal{X}} h(x)e^{\eta^T\varphi(x)}\varphi(x)\left(-e^{-A(\eta)}\right)\nabla A(\eta)^T dx \\
&= \mathbb{E}_{X|\eta}\,\varphi(X)\varphi(X)^T - \left(\mathbb{E}_{X|\eta}\,\varphi(X)\right)\left(\mathbb{E}_{X|\eta}\,\varphi(X)\right)^T \\
&= var_{X|\eta}\,\varphi(X)
\end{aligned}$$

$\nabla^2 A(\eta)$ est donc une matrice semi-définie positive et A est convexe. Dans le cas où A est strictement convexe, la fonction :

$$\begin{array}{rl} \eta \longmapsto & \nabla A(\eta) \\ \mathbb{R}^p \longrightarrow & \mathbb{R}^p \end{array}$$

est injective. On peut alors définir le paramètre de moment μ à partir du paramètre naturel η.

Définition 5.14 *Paramètre de moment*

$$\mu = \nabla A(\eta) = \mathbb{E}_{X|\eta}\, \varphi(X) = \mu(\eta)$$

Exemple 5.5.4 *Loi de Bernouilli(π) :*

$$\frac{dA}{d\eta} = \frac{e^{\eta}}{1+e^{\eta}} = \sigma(\eta) = \pi = \mathbb{E}_{X|\eta}\, X$$

Exemple 5.5.5 *Loi gaussienne(μ,σ) dans $\mathbb{R}$:*

$$\eta = \left(\frac{\mu}{\sigma^2},\, -\frac{1}{\sigma^2}\right)^T, \mu = \left(x,\, x^2\right)^T, \quad A(\eta) = \frac{1}{2}\log\left(-2\eta_2\right) - \frac{\eta_1^2}{4\eta_2}$$

$$\frac{\partial A}{\partial \eta_1} = \mu$$
$$\frac{\partial A}{\partial \eta_2} = \sigma^2 + \mu^2 = \mathbb{E}\left[X^2\right]$$

5.5.2 Liens avec le Maximum de Vraisemblance

Soient $x_1, \ldots, x_N$ des données IID dont la loi est une famille exponentielle. Alors la log-vraisemblance s'écrit :

$$\sum_{n=1}^{N} \log\, p(x_n|\eta) = \sum_{n=1}^{N} \left[\log\, h(x_n) + \eta^T \varphi(x_n) - A(\eta)\right]$$
$$\propto N\left[\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(x_n)\right)^T \eta - A(\eta)\right]$$

D'où son gradient :

$$\nabla(\log - vraisemblance(\eta)) = N\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(x_n) - \nabla A(\eta)\right) = N\left(\frac{1}{N}\sum_{n=1}^{N}\varphi(x_n) - \mu(\eta)\right)$$

et la maximum de vraisemblance atteint pour

$$\mu = \frac{1}{N}\sum_{n=1}^{N}\varphi(x_n) = \langle\varphi(x)\rangle$$

ce qui nous donne un estimateur de $\mathbb{E}_{X|\eta}\, \varphi(X)$.

5.5.3 Liens avec le maximum d'entropie

On cherche a déterminer le paramètre naturel η qui maximise l'entropie de la distribution associée p sur le domaine $\{\eta, A(\eta) < \infty\}$, sous la contrainte d'une moyenne fixée $E_{p(x)}\,\varphi(x) = \mu$. Autrement dit, on cherche la v.a X de distribution p telle que :

$$\max_X H(X) \mid E_{p(x)}\,\varphi(x) = \mu$$

ou encore

$$\max_{p(x)} -\sum_{x\in\mathcal{X}} p(x)\,\log\,p(x) \mid \sum_{x\in\mathcal{X}} p(x)\,\varphi(x) = \mu$$

Il s'agit d'un problème d'optimisation convexe sous contrainte.

Proposition 5.15 *Etant donnés les réalisations $x = (x_1, ..., x_N)$ d'une v.a X, φ une statistique suffisante, h une mesure, et la contrainte sur la moyenne $\mu = \hat{\mu}$ (moyenne empirique),*

$$\begin{aligned} p \text{ maximise l'entropie} &\Leftrightarrow \exists\eta,\ p(u) = \tfrac{1}{Z} e^{\eta^T\varphi(u) - A(\eta)} \ / \ E_{p(u)}\,\varphi(u) = \hat{\mu} \\ &\Leftrightarrow \eta \text{ maximise la vraisemblance } p(x|\eta) \end{aligned}$$

5.5.4 Liens avec les modèles graphiques non orientés

Un cas particulier : le modèle d'Ising

Ici on s'intéresse à des variables aléatoires binaires $X_i \in \{0, 1\}$, $i = 1, ..., N$, telles que :

$$p(x) = p(x_1, ..., x_N) = \frac{1}{Z} \prod_{(i,j)\in E} \psi_{ij}(x_i, x_j)$$

avec

$$\psi_{ij}(x_i, x_j) = V_{ij}^{11} x_i x_j + V_{ij}^{10} x_i(1 - x_j) + V_{ij}^{01}(1 - x_i)x_j + V_{ij}^{00}(1 - x_i)(1 - x_j)$$

Alors p peut s'écrire :

$$p(x) = \frac{1}{Z} \prod_{(i,j)\in E} e^{\theta_{ij} x_i x_j} \prod_{i\in V} e^{\theta_i x_i}$$

ce qui correspond à une famille exponentielle.

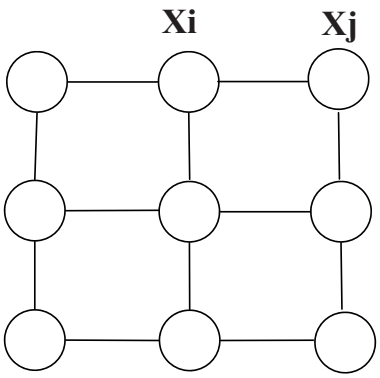

Figure 5.1. Modèle d'Ising.

Cas général

On fait l'hypothèse que p est strictement positive. Alors p s'écrit sous la forme d'une famille exponentielle :

$$\begin{aligned} p(x) &= \frac{1}{Z} \prod_{c \in C} \psi_c(x_c) \\ &= \frac{1}{Z} \prod_{c \in C} \exp\left(\log\, \psi_c(x_c)\right) \\ &= \frac{1}{Z} \exp\left(\sum_{c \in C} \sum_{y_c \in \mathcal{X}_c} \delta(y_c = x_c) \log\, \psi_c(x_c) \right) \end{aligned}$$