Introduction aux modèles graphiques **2007/2008**

Cours 4 — 24 octobre

Enseignant: Francis Bach *Scribe: Mathieu Galtier, Carlos Fernández Granda*

4.1 Estimation de lois à partir de données (Suite)

4.1.1 Estimation de loi gaussienne

Soit une variable $x \in \mathbb{R}$. Nous supposons qu'elle suit une loi normale parametrée par sa moyenne μ et sa variance σ^2 :

$$p(x \mid \mu, \sigma^2) = \frac{e^{\frac{-(x-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \tag{4.1}$$

Soit un échantillon $x_1, \ldots, x_n$ i.i.d. (indépendant et identiquement distribué). La log-vraisemblance est donnée par :

$$\begin{aligned} l\left(\mu, \sigma^2\right) &= \log p\left(x_1, \ldots, x_n \mid \mu, \sigma^2\right) \\ &= \log \prod_{i=1}^{n} p\left(x_i \mid \mu, \sigma^2\right) \\ &= \sum_{i=1}^{n} \left(-\log\left(\sqrt{2\pi}\sigma\right) - \frac{(x_i-\mu)^2}{2\sigma^2} \right) \end{aligned}$$

En dérivant par rapport à μ et σ^2 nous trouvons les estimateurs $\widehat{\mu}$ et $\widehat{\sigma^2}$ qui maximisent la vraisemblance :

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.2}$$

Notons que cette valeur est exactement la moyenne empirique.

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n} \left(x_i - \widehat{\mu}\right)^2}{n} \tag{4.3}$$

Cette valeur est presque la variance empirique (il faudrait changer n par $n-1$ dans le denominateur).

4.1.2 Estimation de loi gaussienne multivariée

Soit une variable $x \in \mathbb{R}^k$. Nous supposons qu'elle suit une loi normale multivariée parametrée par un vecteur de moyennes $\mu \in \mathbb{R}^k$ et une matrice de covariance $\Sigma \in \mathbb{R}^{k \times k}$:

$$p(x \mid \mu, \Sigma) = \frac{e^{\frac{-(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}}}{(2\pi)^{\frac{k}{2}} \sqrt{\det \Sigma}} \tag{4.4}$$

Soit un échantillon $x_1, \ldots, x_n$ i.i.d. . La log-vraisemblance est donnée par :

$$\begin{aligned} l(\mu, \Sigma) &= \log p(x_1, \ldots, x_n \mid \mu, \Sigma) \\ &= \log \prod_{i=1}^{n} p(x_i \mid \mu, \Sigma) \\ &= \sum_{i=1}^{n} \left(-\log (2\pi)^{\frac{k}{2}} - \frac{\log(\det \Sigma)}{2} - \frac{(x_i - \mu)^T \Sigma^{-1} (x_i - \mu)}{2} \right) \end{aligned}$$

Dans ce cas notre fonction est concave. On peut dériver par rapport à μ pour trouver l'estimateur qui maximise la log-vraisemblance. Pour calculer la dérivé nous utiliserons la proposition suivante :

Proposition 4.1 *Soit un vecteur $v \in \mathbb{R}^k$ et une matrice $Q \in \mathbb{R}^{k \times k}$:*

$$\frac{\partial \left(v^T Q v \right)}{\partial v} = 2Qv$$

En appliquant cette proposition :

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

Si on considère que cette expression vaut zéro on trouve l'estimateur du vecteur de moyennes :

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{4.5}$$

Pour calculer l'estimateur de la matrice de covariance d'abord on manipule l'expression de la log-vraisemblance pour faciliter les opérations. On notera $\Lambda = \Sigma^{-1}$.

$$l(\mu, \Sigma) = const + \frac{n \log \det \Lambda}{2} - \frac{\sum_{i=1}^{n} (x - \mu)^T \Sigma^{-1} (x - \mu)}{2}$$

Le dernier terme peut être traité de la façon suivante :

$$\begin{aligned} \frac{\sum_{i=1}^{n} \left((x - \mu)^T \Lambda (x - \mu) \right)}{2} &= \frac{\sum_{i=1}^{n} tr \left((x - \mu)^T \Lambda (x - \mu) \right)}{2} \\ &= \frac{\sum_{i=1}^{n} tr \left(\Lambda (x - \mu) (x - \mu) \right)^T}{2} \\ &= \frac{tr \left(\Lambda \sum_{i=1}^{n} \left((x - \mu) (x - \mu) \right) \right)^T}{2} \end{aligned}$$

Définition 4.2 (Matrice de covariance empirique)

$$\widehat{\Sigma} = \frac{\sum_{i=1}^{n}\left((x-\mu)(x-\mu)^T\right)}{n}$$

L'expression de la log-vraisemblance devient alors :

$$l(\mu,\Lambda) = const + \frac{n\log\det\Lambda}{2} - \frac{n\times tr\left(\Lambda\widehat{\Sigma}\right)}{2}$$

La fonction est la somme d'une fonction concave et d'une fonction linéaire, elle est donc concave. On pourrait essayer de dériver par rapport à chaque élement Λ_{ij}. Mais il est plus pratique de dériver par rapport à toute la matrice :

$$\nabla l(\mu,\Lambda) = \frac{n\Lambda^{-1}}{2} - \frac{n\widehat{\Sigma}}{2}$$

Si on égale cette expression à zéro on trouve :

$$\Lambda^{-1} = \widehat{\Sigma}$$

L'estimateur de la matrice de covariance est alors la matrice de covariance empirique.

(1) ne jamais essayer de dériver par rapport à chaque élément de la matrice Σ ou Λ ;
(2) toujours vérifier dans les produits matriciels que les dimensions sont compatibles.

4.2 Régression linéaire

On modélise le rapport entre une variable $x \in \mathbb{R}^k$ et une variable $y \in \mathbb{R}$. On notera x^i chaque composante de x. On suppose que la probabilité de y conditionnée à x est de la forme :

$$p(y \mid x) = N(\theta^T x, \sigma^2) \tag{4.6}$$

On prend en compte aussi les dépendances où la moyenne de la distribution gaussienne est de la forme $\theta^T x + \theta_0$. Il faut juste redéfinir x comme $\widetilde{x} = (x, 1) \in \mathbb{R}^{k+1}$.
Les données qu'ont utilise pour estimer les paramètres sont de la forme $(x_i^1 \ldots x_i^q, y_i)$ avec $i = 1 \ldots n$, $x_i^j \in \mathbb{R}$ et $y_i \in \mathbb{R}$. Il s'agit de données i.i.d. par paires :

$$\left(x_i^1 \ldots x_i^q, y_i\right) \perp \left(x_j^1 \ldots x_j^q, y_j\right) i \neq j \tag{4.7}$$

On utilise encore la fonction de log-vraisamblance pour trouver une estimation des paramètres :

$$\begin{aligned} l\left(\theta,\sigma^2\right) &= \sum_{i=1}^{n}\left(-\log\left(\sqrt{2\pi}\right) - \log\sigma - \frac{\left(y_i-\theta^T x_i\right)^2}{2\sigma^2}\right) \\ &= const - \frac{n\log\sigma^2}{2} - \sum_{i=1}^{n}\frac{\left(y_i-\theta^T x_i\right)^2}{2\sigma^2} \\ &= const - \frac{n\log\sigma^2}{2} - \frac{\|y-X\theta\|^2}{2\sigma^2} \end{aligned}$$

$X \in \mathbb{R}^{n \times q}$ est une matrice dont la ligne k est de la forme $(x_k^1 \dots x_k^q)$. Pour dériver par rapport à θ on utilisera la proposition suivante :

Proposition 4.3 *Soit un vecteur $v \in \mathbb{R}^k$ et une matrice $Q \in \mathbb{R}^{k \times k}$:*

$$\frac{\partial \left((Qv)^T (Qv) \right)}{\partial v} = 2QQ^T v$$

On obtient alors :

$$\frac{\partial \left(l\left(\theta, \sigma^2\right) \right)}{\partial \theta} = -\frac{X^T \left(X\theta - y \right)}{2\sigma^2}$$

Si cette expression vaut zéro on obtient les **Equations Normales** :

$$X^T X \theta = X^T Y \tag{4.8}$$

L'estimateur pour θ est alors :

$$\widehat{\theta} = \left(X^T X \right)^{-1} X^T Y \tag{4.9}$$

L'estimateur pour σ^2 est :

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} \left(y_i - \widehat{\theta}^T x_i \right)^2}{n} \tag{4.10}$$

Le chapitre 6 du livre décrit l'estimation en ligne de θ.

4.3 Classification Linéaire

Ici nous considérons le cas où les sorties prennent leurs valeurs parmis un nombre fini de possibilités : $Y \in \{1, .., q\}$ et où les entrées sont vectorielles ie : $X = (X^1, \dots, X^k) \in \mathbb{R}^k$. Cette situation s'apparente à un problème de classification.
On pourrait appliquer la méthode de régression linéaire décrite dans la section précedente, mais on trouve deux problèmes :

- La densité parametrée qu'on obtient n'est pas restreinte sur les points y_i. Il faut particulariser la distribution continue en ces points.
- La minimisation de l'écart quadratique pénalise les classifications qui seraient parfaites. La distance $\left\| y_i - \theta^T x \right\|$ peut être grande, tandis que x correspond à la classe i.

Il faut alors utiliser d'autres méthodes. Il y a deux méthodes principales pour approcher ce problème : la méthode discriminative et la méthode générative.

4.3.1 Méthode discriminative

Pour cette méthode on suppose qu'on connait explicitement une expression de $p(y|x, \theta)$, ie on connait f telle que $p(y|x, \theta) = f(x, \theta)$. La méthode du maximum de vraisemblance permet de trouver un estimateur de θ de manière à ce que notre modèle colle au modèle prédictif : le but n'est pas de modéliser à la fois x et y mais de modéliser x sachant y. Ceci est à contraster avec la méthode dite générative.

4.3.2 Méthode générative

Ici nous n'avons pas de d'expression explicite de $p(y|x)$. On définit une loi jointe $p(x, y)$ à partir de laquelle on peut calculer $p(y|x)$. Afin d'estimer les paramêtres, la méthode générative va maximiser la vraisemblance jointe $p(x, y)$ par rapport aux paramêtres (au lieu de maximiser la vraisemblance conditionnelle dans le cas générative).

Le modèle est le suivant :

– Y suit une loi multinomiale (de vecteur Π)

– $\forall j \in \{1, \ldots, q\} \quad X|Y = j$ suit une loi normale : $\mathcal{N}(\mu_j, \Sigma_j)$

Le but du jeu est donc de trouver les μ_j, Σ_j et Π tels que l'on colle au mieux possible à nos données : les couples (x_i, y_i).

D'après le règle de Bayes on a donc

$$\begin{aligned} p(y = i|x) &\propto p(x|y = i) \times p(y = i) \\ &\propto \frac{1}{(2\pi)^{k/2}} \frac{1}{det(\Sigma_i)^{1/2}} exp\Big(-\frac{1}{2}(x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) \Big) \times \Pi_i \\ &\propto \frac{1}{det(\Sigma_i)^{1/2}} exp\Big(-\frac{1}{2} x^T \Sigma_i^{-1} x - \frac{1}{2}\mu_i^T \Sigma_i^{-1}\mu_i + \mu_i^T \Sigma_i^{-1} x \Big) \times \Pi_i \end{aligned}$$

Si on fait l'hypothèse supplémentaire correspondant à LDA (**Linear Discriminant Analysis**) , i.e., $\forall i \in \{1, \ldots, q\} \quad \Sigma_i = \Sigma$, on a :

$$\begin{aligned} p(y = i|x) &\propto \Pi_i \exp\big(-\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i \big) \exp\big(x^T \Sigma^{-1}\mu_i \big) \\ &\propto \exp(b_i) \exp(x^T \theta_i) \end{aligned}$$

En considérant que $b_i = -\frac{1}{2}\mu_i^T \Sigma^{-1}\mu_i + \log(\Pi_i)$ et $\theta_i = \Sigma^{-1}\mu_i$.

En renormalisant on obtient que

$$p(y = i|x) = \frac{e^{x^T \theta_i + b_i}}{\sum_{j=1}^{q} e^{x^T \theta_j + b_j}} \tag{4.11}$$

Définition 4.4 (fonction softmax) *La fonction softmax allant de $\mathbb{R}^q$ dans $\mathbb{R}^q$ associe à chaque vecteur $(z_1, \ldots, z_q)$ le vecteur $S(z_1, \ldots, z_q)$ dont la ième composante s'écrit : $S(z_1, \ldots, z_q)_i = \frac{e^{z_i}}{\sum_{j=1}^{q} e^{z_j}}$. On a bien : $\sum_{j=1}^{q} S(z_1, \ldots, z_q) = 1$ et $S(z_1, \ldots, z_q) \geq 0$*

Avec ce formalisme on peut dire que :

$$p(y = i|x) = S(x^T \theta_1 + b_1, \ldots, x^T \theta_q + b_q)_i \tag{4.12}$$

Si on considère le cas où $\Sigma_i \neq \Sigma_j \quad i \neq j$, nommé QDA (Quadratic Discriminant Analysis), alors les termes quadratiques ne s'annulent pas. Voir DM)

Afin d'estimer les paramêtres, la maximization de la vraisemblance jointe permet d'estimer les paramêtres μ_j, Π_j et Σ (voir DM) qui permettent alors de calculer θ_j et b_j. Dans le cadre discriminatif, les paramêtres θ_j et b_j sont directement estimés.

4.3.3 Regression logistiques : le cas q=2

On peut de la même façon considérer que $Y \in \{0,1\}$, $Y \in \{1,-1\}$ ou $Y \in \{1,2\}$, le choix dépend de ce qu'on veut faire. Pour ce cours nous considèrerons que $Y \in \{0,1\}$.
On a donc d'après ce qui précède :

$$\begin{aligned} p(y=1|x) &= \frac{e^{x^T\theta_1+b_1}}{e^{x^T\theta_0+b_0}+e^{x^T\theta_1+b_1}} \\ &= \frac{1}{1+e^{-\left(x^T(\theta_1-\theta_0)+b_1-b_0\right)}} \\ &= \sigma\big(x^T(\theta_1-\theta_0)+b_1-b_0\big) \end{aligned}$$

où σ est la fonction sigmoïde ie : $\sigma(z) = \frac{1}{1+e^{-z}}$.

En utilisant la même astuce que pour la régression linéaire, on peut s'affranchir du terme constant et considérer le modèle : $p(Y=1|x,\theta) = \sigma(\theta^T x)$.

Calculons désormais la log-vraisemblance afin de la maximiser pour obtenir l'estimateur $\hat{\theta}$:

$$\begin{aligned} l(\theta) &= \sum_i \log\big(p(y=y_i|x_i,\theta)\big) \\ &= \sum_i \log\Big(p(y=1|x_i,\theta)^{\delta_{y_i=1}} p(y=0|x_i,\theta)^{\delta_{y_i=0}}\Big) \\ &= \sum_i y_i \log\big(p(y=1|x_i,\theta)\big) + (1-y_i)\log\big(1-p(y=1|x_i,\theta)\big) \end{aligned}$$

Il vient :

$$l(\theta) = \sum_i y_i \log\sigma(\theta^T x_i) + (1-y_i)\log\big(1-\sigma(\theta^T x_i)\big) \tag{4.13}$$

On note maintenant que $\log(\sigma(z)) = \log\big(\frac{1}{1+e^{-z}}\big) = -\log(1+e^{-z})$ et $1-\sigma(z) = \sigma(-z)$ sont concaves. Donc la log vraisemblance est concave et on peut lui trouver un maximum. Bien qu'il n'y ai pas de formule analytique pour exprimer ce max on peut utiliser des méthodes numériques d'approximation du maximum pour s'en rapprocher.

Méthode de Newton-Raphson La méthode de Newton-Raphson est une méthode itérative, ie telle que x_{t+1} est fonction de x_t, qui vise à trouver l'argmax d'un fonction f.
On fait un développement de Taylor autour de x_t, ce qui donne :

$$f(x) = \underbrace{f(x_t) + \frac{\partial f}{\partial x}(x_t)\big(x-x_t\big) + \frac{1}{2}\big(x-x_t\big)^T \frac{\partial^2 f}{\partial x \partial x^T}(x_t)\big(x-x_t\big)}_{\hat{f}_t(x)} + o\big(||x-x_t||^2\big)$$

Le principe de la méthode de Newton-Raphson est de déterminer x_{t+1} à partir de $\hat{f}_t$. Plus précisément on définit x_{t+1} comme suit : $x_{t+1} = \min_x \hat{f}_t(x)$.
Or

$$\begin{aligned} & \frac{\partial \hat{f}_t}{\partial x} = \frac{\partial f}{\partial x}(x_t) + \frac{\partial^2 f}{\partial x \partial x^T}(x_t)\big(x - x_t\big) = 0 \\ \Leftrightarrow \quad & -\left(\frac{\partial^2 f}{\partial x \partial x^T}(x_t)\right)^{-1} \frac{\partial f}{\partial x}(x_t) = x - x_t \\ \Leftrightarrow \quad & x = x_t - \left(\frac{\partial^2 f}{\partial x \partial x^T}(x_t)\right)^{-1} \frac{\partial f}{\partial x}(x_t) \end{aligned}$$

Définition 4.5 (Algorithme de Newton-Raphson) – *x_0 est quelconque*
– *on passe du rang t à t+1 grâce à l'itération de Newton :*

$$x_{t+1} = x_t - \left(\frac{\partial^2 f}{\partial x \partial x^T}(x_t)\right)^{-1} \frac{\partial f}{\partial x}(x_t)$$

On notera que dans le cas général il peut y avoir différents problèmes lors du déroulement de l'agorithme de Newton-Raphson, par exemple si la hessienne n'est pas inversible ; Par ailleurs dans certains cas on observe des oscillations autour de la valeur cherchée. En revanche "tout se passe bien" dans le cas d'une fonction concave.

Théorème 4.6 *L'algorithme de Newton est globalement convergent pour $l(\theta)$.*

Pour arriver à une approximation convenables on peut raisonnablement faire une trentaine d'itérations.

Application de la méthode de Newton-Raphson à la régression logistique Remarquons dans un premier temps que la dérivée de la fonction sigmoïde est :

$$\sigma'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = \big(1 - \sigma(z)\big)\sigma(z)$$

Ainsi on peut dériver $l(\theta) = \sum_i y_i \log \sigma(\theta^T x_i) + (1 - y_i) \log\big(1 - \sigma(\theta^T x_i)\big)$:

$$\begin{aligned} \frac{dl}{d\theta}(\theta) &= \sum_i y_i \frac{\sigma(\theta^T x_i)}{\sigma(\theta^T x_i)}\big(1 - \sigma(\theta^T x_i)\big)x_i \\ &+ \sum_i (1 - y_i)\frac{-\sigma(\theta^T x_i)}{1 - \sigma(\theta^T x_i)}\big(1 - \sigma(\theta^T x_i)\big)x_i \\ &= \sum_{i=1}^{n} \big(y_i - \sigma(\theta^T x_i)\big)x_i \end{aligned}$$

On définit μ tel que $\forall i \in \{1, \ldots, n\} \quad \mu_i = \sigma(\theta^T x_i)$. Alors on a :

$$\begin{aligned} \frac{dl}{d\theta}(\theta) &= \sum_{i=1}^{n}(y_i - \mu_i)x_i \\ &= X^T(y - \mu) \end{aligned}$$

Calculons maintenant :

$$\begin{aligned} \frac{\partial^2 l}{\partial\theta\partial\theta^T}(\theta) &= -\sum_{i=1}^{n} x_i \big(\frac{\partial\mu_i}{\partial\theta}\big)^T \\ &= -\sum_{i=1}^{n} x_i\mu_i(1-\mu_i)x_i x_i^T \end{aligned}$$

On définit $W = diag\big(\mu_i(1-\mu_i)\big)$. Ainsi

$$\frac{\partial^2 l}{\partial\theta\partial\theta^T}(\theta) = -X^T W X$$

Définition 4.7 (algorithme IRLS) *Dans le cas précis de régression logistique l'algorithme de Newton-Raphson est appelé* ***IRLS*** *(Iterative Reweighted Least Square). Il est implémenté comme suit :*

- $\theta_0 = 0$
- $\theta_{t+1} = \theta_t + (X^T W X)^{-1} X^T (y - \mu)$
- *Critère d'arrèt : pour ϵ petit, en pratique de l'ordre de 10^{-12}.*

$$\begin{aligned} ||\theta_{t+1} - \theta_t|| &< \epsilon \\ ||l(\theta_{t+1}) - l(\theta_t)|| &< \epsilon \end{aligned}$$