Modèles graphiques probabiliste - DM3

Olivier Jais-Nielsen

23 décembre 2010

1 Propagation de messages dans les modeles gaussiens

1.1 Calcul des messages

Comme les variables assoicées aux noeuds suivent une loi Gaussienne scalaire, la loi jointe de toutes ces variables est Gaussiène multivariée :

$$p(x_1,...,x_n) = \exp\left(\eta^T x - \frac{1}{2}x^T \Lambda x - A(\eta)\right)$$

avec

$$A(\eta) = -\frac{1}{2}\left(n\log(2\pi) - \log|\Lambda| + \eta^T \Lambda \eta\right)$$

On a donc, avec $a = \exp(-A(\eta))$,

$$p(x_1,...,x_n) = a\exp\left(\sum_{i\in V}\eta_i x_i - \frac{1}{2}\sum_{i,j\in V} x_i \lambda_{i,j} x_j\right)$$

$$p(x_1,...,x_n) = a\exp\left(\sum_{i\in V}\left(\eta_i x_i - \frac{1}{2}\lambda_{i,i}x_i^2\right) - \frac{1}{2}\sum_{i,j\in E}\lambda_{i,j}x_i x_j\right)$$

car dans un arbre non orienté, l'ensemble des cliques est $V \cup E$.

On peut donc prendre pour potentiels

$$\Psi_i(x_i) = b\exp\left(\eta_i x_i - \frac{1}{2}\lambda_{i,i}x_i^2\right)$$

avec $b = a^{\frac{1}{n}}$ et, pour tout $(i,j) \in E$,

$$\Psi_{i,j}(x_i,x_j) = \exp\left(-\frac{1}{2}\lambda_{i,j}x_i x_j\right)$$

Par définition, le message de i vers j est défini comme

$$m_{i,j}(x_j) = \int_{\mathbb{R}} \Psi_i(x_i)\Psi_{i,j}(x_i,x_j)\prod_{k\in\mathcal{N}(i)-j} m_{k,i}(x_i)\,dx_i$$

Comme marginaliser un vecteur gaussien par rapport à l'une de ses composantes donne lieu à un autre vecteur gaussien, et comme le produit de deux densités gaussiennes et proportionnel à une densité gaussienne, on en déduit que tous les messages sont proportionnels à des densités gaussiennes, c'est à dire des exponentielles de fonctions quadratiques.

On peut donc passer comme messages les espérances $\mu^{(i,j)}$ et les "matrices" de précision $\Lambda^{(i,j)}$ de ces distributions (les messages n'ayant pas besoin d'être normalisés). On a alors

$$\begin{cases} \Lambda^{(i,j)} = & -\lambda_{i,j}^2 \left(\lambda_{i,i} - \sum_{k \in \mathscr{N}(i)-j} \Lambda^{(k,i)} \right)^{-1} \\ \mu^{(i,j)} = & \frac{1}{\lambda_{i,j}} \left(\eta_i + \sum_{k \in \mathscr{N}(i)-j} \Lambda^{(k,i)} \mu^{(k,i)} \right) \end{cases}$$

1.2 Marginales

Le calcul des marginales au niveau des noeuds donne les espérances marginales :

$$E(X_i) = \mu^{(i)} = \left(\eta_i + \sum_{k \in \mathscr{N}(i)} \Lambda^{(k,i)} \mu^{(k,i)} \right) \left(\lambda_{i,i} - \sum_{k \in \mathscr{N}(i)} \Lambda^{(k,i)} \right)^{-1}$$

On a de plus :

$$E(X_i X_j) = \frac{1}{\lambda_{i,j}} + \mu^{(i)} \mu^{(j)}$$

1.3 Complexité

On a $2(n-1)$ messages. Le calcul de chaque message est d'une complexité finie indépendante des espaces. Le calcul avec cette méthode de l'espérance à partir de la précision et du vecteur de paramètres canoniques avec cette méthode est donc linéaire.

Dans le cas général d'un modèle Gaussien, l'opération à faire est résoudre en μ le système $\Lambda\mu = \eta$, système linéaire symétrique qui se résoud donc en une complexité de n^2. Dans le cas d'un arbre, la méthode précédente est d'un ordre de complexité plus rapide.

2 Apprentissage de la structure d'un arbre

2.1 Préliminaire 1

On a

$$p(x^n|\eta) = \prod_x p(x|\eta)^{\delta(x^n=x)}$$

donc

$$p(x^n|\eta) = \prod_x \eta(x)^{\delta(x^n=x)}$$

et donc la log-vraisemblance conditionnelle s'écrit

$$\ell(\eta) = \sum_{n=1}^{N} \sum_x \delta(x^n = x) \log(\eta(x))$$

et finalement

$$\ell(\eta) = \sum_x N\hat{p}(x)\log(\eta(x))$$

On cherche à la maximiser en fonction de η avec $\eta \geq 0$ et $\sum_x \eta(x) = 1$. On considère le Lagrangien suivant

$$\mathscr{L}(\eta,\lambda) = \ell(\eta) + \lambda\left(\sum_x \eta(x) - 1\right)$$

Il s'agit d'une fonction concave qui assure bien les contraintes (la contrainte de positivité de η est assurée par l'expression de $\ell(\eta)$ qui explose lorsque $\eta \to 0^+$). On la maximise par différentiation et on obtient

$$\begin{cases} \frac{\partial \mathscr{L}}{\partial \lambda} = \sum_x \eta(x) - 1 = 0 \\ \frac{\partial \mathscr{L}}{\partial \eta(x)} = N\frac{\hat{p}(x)}{\eta(x)} + \lambda = 0 \quad \forall x \end{cases}$$

On a alors bien la contrainte d'égalité $\sum_x \eta(x) = 1$ et de plus

$$\lambda\eta(x) = -N\hat{p}(x)$$

soit en sommant sur x

$$\lambda = -N\sum_x \hat{p}(x) = N$$

et donc

$$\eta^*(x) = \hat{p}(x)$$

La log-vraisemblance conditionnelle maximale est donc

$$\ell^* = \ell(\eta^*) = \sum_x N\hat{p}(x)\log(\hat{p}(x)) = -NH(X)$$

2.2 Préliminaire 2

On a

$$p(y^n|x^n,\eta) = \prod_{x,y} p(y|y,\eta)^{\delta(y^n=y,x^n=x)}$$

donc

$$p(y^n|x^n,\eta) = \prod_{x,y} \eta(x,y)^{\delta(y^n=y)\delta(x^n=x)}$$

et donc la log-vraisemblance conditionnelle s'écrit

$$\ell(\eta) = \sum_{n=1}^{N}\sum_{x,y} \delta(x^n = x)\,\delta(y^n = y)\log(\eta(x,y))$$

et finalement

$$\ell(\eta) = \sum_{x,y} N\hat{p}(x,y)\log(\eta(x,y))$$

On cherche à la maximiser en fonction de η avec $\eta \geq 0$ et $\sum_y \eta(x,y) = 1$ pour tout x. On considère le Lagrangien suivant

$$\mathscr{L}(\eta,\lambda) = \ell(\eta) + \sum_x \lambda_x \left(\sum_y \eta(x,y) - 1 \right)$$

Il s'agit d'une fonction concave qui assure bien les contraintes (la contrainte de positivité de η est assurée par l'expression de $\ell(\eta)$ qui explose lorsque $\eta \to 0^+$). On la maximise par différentiation et on obtient

$$\begin{cases} \frac{\partial \mathscr{L}}{\partial \lambda_x} = \sum_y \eta(x,y) - 1 = 0 & \forall x \\ \frac{\partial \mathscr{L}}{\partial \eta(x)} = N \frac{\hat{p}(x,y)}{\eta(x,y)} + \lambda_x = 0 & \forall x \end{cases}$$

On a alors bien la contrainte d'égalité $\sum_y \eta(x,y) = 1$ pour tout x et de plus

$$\lambda_x \eta(x,y) = -N\hat{p}(x,y)$$

soit en sommant sur y

$$\lambda_x = -N \sum_y \hat{p}(x,y) = -N\hat{p}(x)$$

et donc

$$\eta^*(x,y) = \frac{\hat{p}(x,y)}{\hat{p}(x)}$$

La log-vraisemblance conditionnelle maximale est donc

$$\ell^* = \ell(\eta^*) = \sum_{x,y} N\hat{p}(x,y) \log \left(\frac{\hat{p}(x,y)}{\hat{p}(x)} \right) = N(H(X) - H(X,Y))$$

2.3 Log-vraisemblance maximale pour un arbre couvrant

Dans un arbre couvrant T, chaque noeud à au plus un parent. On note, pour tout noeud d'indice p l'indice de son parent $\pi_p(T)$, et on pose $\pi_p(T) = \phi$ si le noeud n'en a pas. En utilisant la convention $p\left(x_p | x_\phi\right) = p(x_p)$, alors toute loi jointe se factorisant sur l'arbre s'écrit :

$$p(x_1,...,x_p) = \prod_{p=1}^{P} p\left(x_p | x_{\pi_p(T)}\right)$$

La paramétrisation la plus générale est donc données par $\eta = (\eta_1,...,\eta_p)$ ou η_p : $x \mapsto p(x_p)$ si $\pi_p = \phi$ et $\eta_p : x,y \mapsto p\left(X_p = x | X_{\pi_p(T)} = y\right)$.

La log-vraisemblance s'écrit

$$\ell(T,\eta) = \log\left(p\left(\left(x_1^1,...,x_P^1\right),...,\left(x_1^N,...,x_P^N\right)|\eta\right)\right) = \sum_{n=1}^{N} \log\left(p\left(x_1^n,...,x_P^n|\eta\right)\right)$$

mais

$$p(x_1^n,...,x_P^n|\eta) = \prod_{p=1}^{P} p\left(x_p^n | x_{\pi_p(T)}^n, \eta\right) = \prod_{p=1}^{P} p\left(x_p^n | x_{\pi_p(T)}^n, \eta_p\right)$$

et donc

$$\ell(T,\eta) = \sum_{n=1}^{N}\sum_{p=1}^{P} \log\left(p\left(x_p^n | x_{\pi_p(T)}^n, \eta_p\right)\right) = \sum_{p=1}^{P} \ell_p(\eta_p)$$

avec

$$\ell_p(\eta_p) = \sum_{n=1}^{N} \log\left(p\left(x_p^n | x_{\pi_p(T)}^n, \eta_p\right)\right)$$

Or d'après la question précédente, la fonction ℓ_p a pour valeur maximale

$$\ell_p^* = -NH(X_p)$$

si $\pi_p(T) = \phi$ et

$$\ell_p^* = N\left(H\left(X_{\pi_p(T)}\right) - H\left(X_{\pi_p(T)}, X_p\right)\right)$$

dans le cas contraire. Avec les conventions choisies, la fonction $\ell(T,.)$ a donc bien pour valeur maximale

$$\ell(T) = \sum_{p=1}^{P} \ell_p^* = N\sum_{p=1}^{P}\left(H\left(X_{\pi_p(T)}\right) - H\left(X_{\pi_p(T)}, X_p\right)\right)$$

2.4 Information mutuelle empirique

On considère la divergence de Kullback-Leiber D des lois $p(X_p, X_q)$ et $p(X_p)\,p(X_q)$. On a

$$\begin{aligned} D &= D(p(X_p,X_q)\,||\,p(X_p)\,p(X_q)) \\ &= \sum_{x_p,x_q} p(x_p,x_q)\log\frac{p(x_p,x_q)}{p(x_p)\,p(x_q)} \\ D &= \sum_{x_p,x_q} p(x_p,x_q)\log(p(x_p,x_q)) - \sum_{x_p,x_q} p(x_p,x_q)\log(p(x_p)) - \sum_{x_p,x_q} p(x_p,x_q)\log(p(x_q)) \end{aligned}$$

mais

$$\sum_{x_p,x_q} p(x_p,x_q)\log(p(x_p)) = \sum_{x_p}\log(p(x_p))\sum_{x_q} p(x_p,x_q) = \sum_{x_p}\log(p(x_p))\,p(x_p)$$

et donc

$$D = \sum_{x_p,x_q} p(x_p,x_q)\log(p(x_p,x_q)) - \sum_{x_p} p(x_p)\log(p(x_p)) - \sum_{x_q} p(x_q)\log(p(x_q))$$

On a donc

$$I(X_p,X_q) = D(\hat{p}(X_p,X_q)\,||\,\hat{p}(X_p)\,\hat{p}(X_q))$$

Comme la divergence de Kullback-Leiber est positive ou nulle, on en déduit que l'information mutuelle est également positive ou nulle.

2.5 Arbre couvrant orienté de vraisemblance maximale

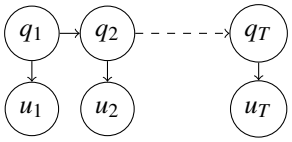
On a

$$\ell(T) = N \sum_{p=1}^{P} \left(I\left(X_p, X_{\pi_p(T)}\right) - H(X_p) \right) = N \sum_{p=1}^{P} I\left(X_p, X_{\pi_p(T)}\right) - N \sum_{p=1}^{P} H(X_p)$$

Maximiser ℓ par rapport à T revient donc à maximiser $\sum_{p=1}^{P} I\left(X_p, X_{\pi_p(T)}\right)$ qui est une somme de termes positifs. Cela revient donc à maximiser $I\left(X_p, X_{\pi_p(T)}\right)$ pour tout noeud p. Ce problème est un problème de recherche d'arbre couvrant de poids minimal (ici, le poids doit être décroissant avec I). En effet, on considère le graphe non orienté totalement connecté ; le but est de trouver un arbre couvrant tous les sommets et maximisant l'information mutuelle des arêtes.

3 Implémentation - HMM

Le modèle graphique étudié est le suivant :



avec

$$u_t = (x_t, y_t)$$

et

$$p(q_{t+1} = j | q_t = i) = a_{i,j}$$

et

$$p(q_t = i) = \pi_i$$

et

$$p(u_t | q_t = i) = \frac{1}{2\pi\sqrt{\det \Sigma_i}} \exp\left(-\frac{1}{2} (u_t - \mu_i)^T \Sigma_i^{-1} (u_t - \mu_i) \right)$$

3.1 Récursions α et β

On calcul les récursions en log pour éviter les problèmes d'underflow :

$$\begin{cases} \log \alpha_1(q_1) = \log p(u_1 | q_1) + \log p(q_1) \\ \log \alpha_{t+1}(q_{t+1}) = \log(u_{t+1} | q_{t+1}) + \log \sum_{q_t} \exp(\log p(q_{t+1} | q_t) + \log \alpha_t(q_t)) \end{cases}$$

et

$$\begin{cases} \log \beta_T(q_T) = 0 \\ \log \beta_t(q_t) = \log \sum_{q_{t+1}} \exp(\log p(q_{t+1} | q_t) + \log \beta_{t+1}(q_{t+1}) + \log(u_{t+1} | q_{t+1})) \end{cases}$$

avec

$$\log\left(u_{t+1}|q_{t+1}=i\right)=-\frac{1}{2}\left(\log\left(2\pi\right)+\log|\Sigma_i|-\left(u_t-\mu_i\right)^T\Sigma_i^{-1}\left(u_t-\mu_i\right)\right)$$

et

$$\log p\left(q_{t+1}=j|q_t=i\right)=\log\left(a_{i,j}\right)$$

et

$$\log p\left(q_t=i\right)=\log\left(\pi_i\right)$$

De plus, les logarithmes de sommes sont calculés de la façon suivante :

$$\log\sum_{x_i}\exp\left(x_i\right)=X+\log\left(1+\sum_{x_i\neq X}\exp\left(x_i-X\right)\right)$$

avec $X=max_i\left(x_i\right)$.

On a alors

$$\log p\left(q_t,u_1,...,u_T\right)=\log\alpha_t\left(q_t\right)+\log\beta_t\left(q_t\right)$$

et

$$\log p\left(q_t|u_1,...,u_T\right)=\log p\left(q_t,u_1,...,u_T\right)-\log\sum_{u_1,...,u_T}\exp\left(\log p\left(q_t,u_1,...,u_T\right)\right)$$

De même

$$\log p\left(q_t,q_{t+1},u_1,...,u_T\right)=\log p\left(q_{t+1}|q_t\right)+\log\alpha_t\left(q_t\right)+\log\beta_{t+1}\left(q_{t+1}\right)+\log p\left(u_{t+1}|q_{t+1}\right)$$

et

$$\log p\left(q_t,q_{t+1}|u_1,...,u_T\right)=\log p\left(q_t,q_{t+1},u_1,...,u_T\right)-\log\sum_{u_1,...,u_T}\exp\left(\log p\left(q_t,q_{t+1},u_1,...,u_T\right)\right)$$

3.2 Equations d'estimation de EM

Dans le cas d'un HMM, la log-vraisemblance complète s'écrit

$$\ell_c=\sum_q\delta\left(q_1=q\right)+\sum_t\sum_{q_1,q_2}\delta\left(q_t=q_1\right)\delta\left(q_{t+1}=q_2\right)\log a_{q_1,q_2}+\sum_t\sum_q\delta\left(q_t=q\right)\log p\left(u_t|q_t=q\right)$$

Pour maximiser ℓ_c, on doit avoir$\pi_i=\delta\left(q_1=i\right)$ et $a_{i,j}=\frac{m_{i,j}}{\sum_j m_{i,j}}$ avec $m_{i,j}=\delta\left(q_t=i\right)\delta\left(q_{t+1}=j\right)$. On estime alors$\pi$ par

$$\hat{\pi}_i=p\left(q_1=i|u_1,...,u_T\right)$$

et a par

$$\hat{a}_{i,j}=E\left(m_{i,j}|u_1,...,u_T\right)=\sum_t p\left(q_t=i,q_{t+1}=j|u_1,...,u_T\right)$$

et μ par

$$\hat{\mu}_i=\frac{\sum_t p\left(q_t=i|u_t\right)u_t}{\sum_t p\left(q_t=i|u_t\right)}$$

etΣpar

$$\hat{\Sigma}_i = \frac{\sum_t p\left(q_t = i|u_t\right)\left(u_t - \hat{\mu}_i\right)\left(u_t - \hat{\mu}_i\right)^T}{\sum_t p\left(q_t = i|u_t\right)}$$

avec $p\left(q_t|u_t\right) = \frac{p(q_t)p(u_t|q_t)}{\sum_{q_t} p(q_t)p(u_t|q_t)}$