Homework 1 PGM

Sharone Dayan, Michael Sutton

October 2017

1 Learning in Discrete Graphical Model

Considérons le modèle suivant :
Soit Z et X deux variables discrètes qui prennent respectivement M et K valeurs tel que :

$$\begin{aligned} p(Z=m) &= \pi_m \qquad \forall m \in [\![1,\dots,M]\!] \\ p(X=k|Z=m) &= \theta_{mk} \qquad \forall k \in [\![1,\dots,K]\!], \forall m \in [\![1,\dots,M]\!] \end{aligned}$$

Données : On a un echantillon $(z^{(1)},x^{(1)}),\dots,(z^{(N)},x^{(N)})$ d'observations i.i.d du couple (Z,X).

Posons Y $=(Y_1,\dots,Y_M)$ avec $Y_m = \mathbb{1}_{Z=m}$. Posons $B = \begin{pmatrix} B_1 \\ \vdots \\ B_M \end{pmatrix}$ avec $(B_m)_k = \mathbb{1}_{X=k|Z=m}$, on a alors :

$$\begin{aligned} p(Y_m=m) = p(Z=m) &= \pi_m \qquad \forall m \in [\![1,\dots,M]\!] \\ p(Y=y) &= \prod_{m=1}^{M} \pi_m^{y_m} \end{aligned}$$

L'évènement $\{Y=k\}$ correspond à $\{Y_k = 1 \ et \ Y_l = 0 \ \ \forall l \neq k\}$, $Y \in [0,1]^K$.
Et concernant B et Z on a les relations :

$$\begin{aligned} p(B_{mk}=1) = p((B_m)_k=1) = p(X=k|Z=m) &= \theta_{mk} \qquad \forall k \in [\![1,\dots,K]\!], \forall m \in [\![1,\dots,M]\!] \\ p(B=b) &= \prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{mk}^{b_{mk}} \end{aligned}$$

mettre un texte comme pour Y decrivant B

Ecrivons la fonction de vraisemblance :

$$\begin{aligned}
L((\pi,\theta)) &= p_{(\pi,\theta)}((Z^{(1)},X^{(1)})=(z^{(1)},x^{(1)}),\dots,(Z^{(N)},X^{(N)})=(z^{(N)},x^{(N)})) \\
&= \prod_{n=1}^{N} p_{(\pi\theta)}((Z^{(n)},X^{(n)})=(z^{(n)},x^{(n)})) \qquad \text{par independance} \\
&= \prod_{n=1}^{N} p(X^{(n)}=x^{(n)}|Z^{(n)}=z^{(n)})\, p(Z^{(n)}=z^{(n)}) \\
&= \prod_{n=1}^{N} p(B^{(n)}=b^{(n)})\, p(Y^{(n)}=y^{(n)}) \\
&= \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{mk}^{b_{mk}^{(n)}} \prod_{m\prime=1\prime}^{M} \pi_{m\prime}^{y_{m\prime}^{(n)}} \\
&= \prod_{n=1}^{N}\prod_{m=1}^{M}\prod_{k=1}^{K} \theta_{mk}^{b_{mk}^{(n)}} \pi_m^{y_m^{(n)}}
\end{aligned}$$

En remarquant que les cas ou $\pi_i = 0$ (resp. $\theta_{mk} = 0$) entrainerai que $y_i = 0$ sur toute les obesrvation (resp. $B_{mk} = 0$ sur toute les obervations), les valeurs correspondantes dans le produit valent 1 et "disparaissent". On en conclu donc que la log vraissemblance est bien defini et vaut :

$$\begin{aligned}
l((\pi,\theta)) &= \sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{m=1}^{M} b_{mk}^{(n)}\log(\theta_{mk}) + y_m^{(n)}\log(\pi_m) \\
&= \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{n=1}^{N} b_{mk}^{(n)}\log(\theta_{mk}) + y_m^{(n)}\log(\pi_m)
\end{aligned}$$

En notant $n_m = \sum_{n=1}^{N} y_m^{(n)}$, qui correspond au nombre d'observations de Z prenant la valeur j, et en notant $n_{mk} = \sum_{n=1}^{N} b_{mk}^{(n)}$ qui correspond au nombre d'observations de (Z, X) prenant la valeur (m, k) on a :

$$l((\pi,\theta)) = \sum_{m=1}^{M}\sum_{k=1}^{K} n_{mk}\log(\theta_{mk}) + n_m\log(\pi_m)$$

L'objectif est donc de maximiser la fonction de log vraisemblance $l((\pi,\theta))$ sous les contrainte que $\sum_{m=1}^{M}\pi_m = 1$ et $\forall m\ \sum_{k=1}^{K}\theta_{mk} = 1$, ou en d'autres termes :

$$\min_{\substack{\pi_m \geq 0 \\ \theta_{mk}\geq 0}} f((\pi,\theta)) = -l((\pi,\theta)) = \sum_{m=1}^{M}\sum_{k=1}^{K} - n_{mk}\log(\theta_{mk}) - n_m\log(\pi_m) \qquad s.c \quad 1^T\pi = 1 \in \mathbb{R} \ \ and \ \ 1^T\theta = 1 \in \mathbb{R}^M$$

Le Lagrangien de ce problème donne :

$$\mathcal{L}(\pi,\theta,\lambda) = -l((\pi,\theta)) + \lambda_0(\sum_{m=1}^{M}\pi_m - 1) + \sum_{m=1}^{M}(\lambda_j(\sum_{k=1}^{K}\theta_{mk} - 1))$$

Il est évident que $n_i \geq 0$ car $y_i \geq 0$ donc f est convexe comme somme de fonction convexe. De plus l'ensemble $\{\pi_m \geq 0, \theta_{mk} \geq 0\ , \forall m \in [\![1,\dots,M\ \forall k \in [\![1,\dots,K]\!]\}$ est convexe, il s'agit d'un problème d'optimisation convexe. Les contraintes son linéaires, et il existe $\pi_1, \pi_2, ..., \pi_M \;\; tq \;\; \sum_{i=1}^{M} \pi_i = 1$, donc d'après la qualification de contraintes de Slater, le problème a la propriété de forte dualité. Ainsi :

$$\min_{\substack{\pi_m \geq 0 \\ \theta_{mk} \geq 0}} f((\pi, \theta)) = \max_{\lambda} \min_{\substack{\pi_m \geq 0 \\ \theta_{mk} \geq 0}} \mathcal{L}(\pi, \theta, \lambda)$$

Je suis pas sur que c'est phrase soit la bonne justificationComme $L(\pi, \lambda)$ est convexe par rapport à π, le minimum se trouve en annulant la dérivée de $L(\pi, \lambda)$ par rapport à π.
Ainsi, on obtient :

$$\frac{\partial \mathcal{L}}{\partial \pi_i} = -\frac{K\, n_i}{\pi_i} + \lambda = 0 \qquad \forall i \in [\![1, \dots, M]\!]$$

Donc :

$$\pi_i \lambda = K n_i \qquad \forall i \in [\![1, \dots, M]\!]$$

En substituant cette égalité à la contrainte $\sum_{i=1}^{M} \pi_i = 1$, on obtient $\lambda = K \sum_{i=1}^{M} n_i$, d'où $\lambda = kN \; (\neq \;\; 0 \;)$ avec N le nombre d'observations.
On obtient finalement :

$$\widehat{\pi}_i = \frac{K n_i}{\lambda} = \frac{\sum_{j=1}^{n} y_i^{(j)}}{n} = \frac{n_i}{n} \qquad \forall i \in [\![1, \dots, M]\!]$$

Nous cherchons donc à trouver :

$$\widehat{\pi}_{ML} = argmax_{\pi_m} L(z_1, ..., z_M; \pi_m)$$