Homework 2 - Probabilistic Graphical Models

Matthis Maillard

November 2018

1 Exercice 1

1. The proposition $X \perp\!\!\!\perp Y|T$ for any $p(x) \in \mathcal{L}(G)$ is not true. Lets take the example where X and Y are random, binary and independent variables, $Z = X + Y$ and $T = Z$. (X,Y,Z,T) verify the structure of the graph G.

$$P(X, Y|T) = P(X|Y, T)P(Y|T) = P(X|X)P(Y|T) \text{ because } X = T - Y$$

$P(X|X)$ is either equal to 0 or 1. Hence, $P(X|X) \neq P(X|T)$. Thus, $P(X, Y|T) \neq P(X|T)P(Y|T)$: the propostition $X \perp\!\!\!\perp Y|T$ for any $p(x) \in \mathcal{L}(G)$ is false.

2. (a) (X,Y,Z) are real variables on a finite space and Z is a binary variable: $P(Z = 1) = p$ and $P(Z = 0) = 1 - p$ with $p \in [0, 1]$. Let us assume that $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y|Z$. We have that:

$$P(X, Y) = P(X, Y|Z = 0)(1-p)+P(X, Y|Z = 1)p = P(X|Z = 0)P(Y|Z = 0)(1-p)+P(X|Z = 1)P(Y|Z = 1)p$$

$$\text{and } P(X, Y) = P(X)P(Y) = [P(X|Z = 0)(1 - p) + P(X|Z = 1)p][P(Y|Z = 0)(1 - p) + P(Y|Z = 1)p]$$

Thus, by writing the equality between the two formula and factorizing:

$$(1 - p)p[P(X|Z = 0)[P(Y|Z = 0) - P(Y|Z = 1)] + P(X|Z = 1)[P(Y|Z = 1) - P(Y|Z = 0)]] = 0$$

$$[P(Y|Z = 1) - P(Y|Z = 0)][P(X|Z = 1) - P(X|Z = 0)] = 0$$

Hence, either $P(Y|Z = 1) = P(Y|Z = 0)$ or $P(X|Z = 1) = P(X|Z = 0)$, as Z is a binary variable, it means that either $P(Y) = P(Y|Z = 1) = P(Y|Z = 0)$ or $P(X) = P(X|Z = 1) = P(X|Z = 0)$. Finally, we have that either $X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$.

(b) This statement is not true in general. If we take X,Y independent random variables and V,W independent random variables with $X \perp\!\!\!\perp W$, X dependent of V, $Y \perp\!\!\!\perp V$ and Y dependent of W. We write Z=(V,W). We have $X, V \perp\!\!\!\perp Y, W$, $P(X|Z) = \frac{P(X,V,W)}{P(V,W)} = \frac{P(X,V)}{P(V)} = P(X|V)$ and also $P(Y|Z) = P(Y|W)$. Thus,

$$\begin{aligned} P(X, Y|Z) &= \frac{P(X, Y, V, W)}{P(V, W)} \\ &= \frac{P(X, V)P(Y, W)}{P(V)P(W)} \\ &= P(X|V)P(Y|W) \\ &= P(X|Z)P(Y|Z) \end{aligned}$$

So, we have that $X \perp\!\!\!\perp Y$ and $X \perp\!\!\!\perp Y|Z$, yet X and Y are both dependent of Z. We conclude that the statement is false in general.

2 Exercice 2

1. $E' = (E \backslash \{i \to j\}) \bigcup \{j \to i\}$ so: $\pi_i^{G'} = \pi_i^G \bigcup \{j\}$, $\pi_j^{G'} = \pi_i^G$ and for all $k \neq i, j$, $\pi_k^G = \pi_k^{G'}$
Let $p(x) \in \mathcal{L}(G)$.

$$\begin{aligned} p(x) = \prod_{k=1}^{n} p(x_k|x_{\pi_k^G}) &= p(x_i|x_{\pi_i^G})p(x_j|x_{\pi_j^G}) \prod_{k\neq i,j}^{n} p(x_k|x_{\pi_k^G}) \\ &= p(x_i|x_{\pi_i^G})p(x_j|x_{\pi_i^G}, x_i) \prod_{k\neq i,j}^{n} p(x_k|x_{\pi_k^G}) \\ &= p(x_j, x_i|x_{\pi_i^G}) \prod_{k\neq i,j}^{n} p(x_k|x_{\pi_k^G}) \\ &= p(x_i|x_j, x_{\pi_i^G})p(x_j|x_{\pi_i^G}) \prod_{k\neq i,j}^{n} p(x_k|x_{\pi_k^G}) \\ &= p(x_i|x_{\pi_i^{G'}})p(x_j|x_{\pi_j^{G'}}) \prod_{k\neq i,j}^{n} p(x_k|x_{\pi_k^{G'}}) \end{aligned}$$

Hence, $p(x) \in \mathcal{L}(G')$, as with the same reasoning we can show that $\mathcal{L}(G') \subset \mathcal{L}(G)$, we conclude that $\mathcal{L}(G') = \mathcal{L}(G)$

2. First, we show that $\mathcal{L}(G) \subset \mathcal{L}(G')$. Let $p(x) \in \mathcal{L}(G)$, x_r be the root of G and x_{r_c} be one of its children. As G contains no v-structure and is a DAG, the maximal cliques of G' are the edges and except for the root, a node has exactly one parent. We have :

$$p(x) = \prod_{i=1}^{n} p(x_i|x_{\pi_i}) = p(x_r)p(x_{r_c}|x_r) \prod_{\substack{i=1 \\ i\neq r \\ i\neq r_c}}^{n} p(x_i|x_{\pi_i})$$

If we write $\Psi_{(r,r_c)}(x_r, x_{r_c}) = p(x_r)p(x_{r_c}|x_r)$ and for all i, $i \neq r$, $i \neq r_c$ $\Psi_{(i,\pi_i)}(x_i, x_{\pi_i}) = p(x_i|x_{\pi_i})$.
Then, we have that $p(x) = \prod_{e\in E} \Psi_e(x_e)$ where E is the set of the edges and the set of the maximal cliques. By property of a probability, for all $e \in E$, $\Psi_e(x_e) \geq 0$ and $\sum_x p(x) = 1$. Thus, $p(x)$ can be written in the form: $p(x) = \frac{1}{\sum_x \prod_{e\in E} \Psi_e(x_e)} \prod_{e\in E} \Psi_e(x_e)$. We conclude that $p(x) \in \mathcal{L}(G')$, hence $\mathcal{L}(G) \subset \mathcal{L}(G')$.

Let $p(x)$ be in $\mathcal{L}(G')$.

$$p(x) = \frac{1}{Z} \prod_{v\in V} \Psi_v(x_v) \prod_{e\in E} \Psi_e(x_e) \quad \text{where V is the set of vertices of G}$$

for all $e \in E$ there is $i \in \{1, .., n\} \backslash \{r\}$ such that $\Psi_e(x_e) = \Psi_{i,\pi_i}(x_i, x_{\pi_i})$. In fact all i except r are covered by this property. Thus, by writing: for all $i \in \{1, .., n\} \backslash \{r\}$ $f_i(x_i, x_{\pi_i}) = \Psi_i(x_i)\Psi_{i,\pi_i}(x_i, x_{\pi_i})$ and $f_r(x_r, x_{\pi_r}) = \Psi_r(x_r)$ we can rewrite p(x):

$$p(x) = \prod_{i=1}^{n} \frac{1}{Z^{\frac{1}{n}}} f_i(x_i, x_{\pi_i})$$

Hence $p(x) \in \mathcal{L}(G)$. Finally, we have $\mathcal{L}(G) = \mathcal{L}(G')$

3 Exercice 3

1. The different results obtained by k-means algorithm are similar, yet the centers are never exactly the same and the measure of distortion varies from 3237 to 3240. The resulting clusters are also not always the same, especially in the central zone of the figure. The result is very dependant of the initialization.

2. In this problem, we need to maximize the loss: $\mathcal{L}(q,\theta) = \sum_z q(z)\log(\frac{p_\theta(x,z)}{q(z)})$ with $Z \sim \mathcal{M}(\pi_1,..,\pi_K)$, $(X|Z=k) \sim \mathcal{N}(\mu_k, \sigma_k^2 I_d)$ and $\theta = (\pi, \mu, \sigma_k^2 I_d)$. To maximize $\mathcal{L}(q,\theta)$ w.r.t q we have to set $q(z) = p_{\theta t}(z|x)$. Let $(x_i, z_i)_{i\in\{1,..,N\}}$ be an i.i.d sample. For the E-step, we have to compute:

$$p_\theta(z=k|x_i) = \frac{p_\theta(x_i|z=k)p_\theta(z_i=k)}{\sum_{k'} p_\theta(x_i|z=k')p_\theta(z_i=k')} = \frac{\frac{\pi_k}{(2\pi\sigma_k^2)^{\frac{d}{2}}}\exp(\frac{1}{2\sigma_k^2}||x_i-\mu_k||_2^2)}{\sum_{k'}\frac{\pi_{k'}}{(2\pi\sigma_{k'}^2)^{\frac{d}{2}}}\exp(\frac{1}{2\sigma_{k'}^2}||x_i-\mu_{k'}||_2^2)} = \tau_i^k$$

then, with $z_i^k = 1$ if $z_i = k$ and 0 otherwise

$$\begin{aligned}\log(p_\theta(x,z)) &= \log(p_\theta(z)) + \log(p_\theta(x|z)) \\ &= \sum_{i=1}^N\sum_{k=1}^K z_i^k(\log(\pi_k) - \frac{d}{2}\log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}||x_i-\mu_k||_2^2)\end{aligned}$$

By seeing that $\mathbb{E}_{Z|X}(z_i^k) = \tau_i^k$, we conclude that to maximize $\mathcal{L}(q,\theta)$, w.r.t θ we have to maximize

$$\sum_{i=1}^N\sum_{k=1}^K \tau_i^k(\log(\pi_k) - \frac{d}{2}\log(2\pi\sigma_k^2) - \frac{1}{2\sigma_k^2}||x_i-\mu_k||_2^2)$$

As it is concave, to find the maximum, we set the gradients to zero. The values obtained are:

$$\begin{cases} \pi_k = \frac{1}{N}\sum_{i=1}^N \tau_i^k \\ \mu_k = \frac{\sum_{i=1}^N \tau_i^k x_i}{\sum_{i=1}^N \tau_i^k} \\ \sigma_k^2 = \frac{\sum_{i=1}^N \tau_i^k (x_i-\mu_k)^T(x_i-\mu_k)}{d\sum_{i=1}^N \tau_i^k} \end{cases}$$

3. In the general case, only Σ changes. We have:

$$\Sigma_k = \frac{\sum_{i=1}^N \tau_i^k (x_i-\mu_k)(x_i-\mu_k)^T}{\sum_{i=1}^N \tau_i^k}$$

4. The measures of the log-likelihoods are:

	Train set	Test set
isotropic	-2676.6	-2654.2
general	-2344.7	-2425.9

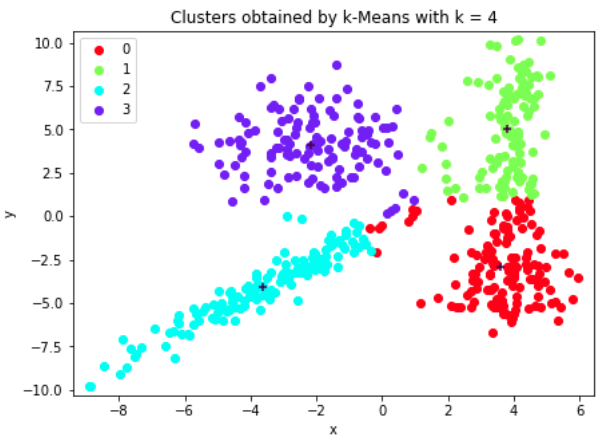
It is normal that the isotropic case has always a lower log-likelihood than the general case because it is just a specific case of the general. The isotropic case assumes that the distribution of the data in a cluster is the same for all directions. We can see that it is not a realistic assumption in the figures bellow. The cluster 3 seems to be stretched in one direction.

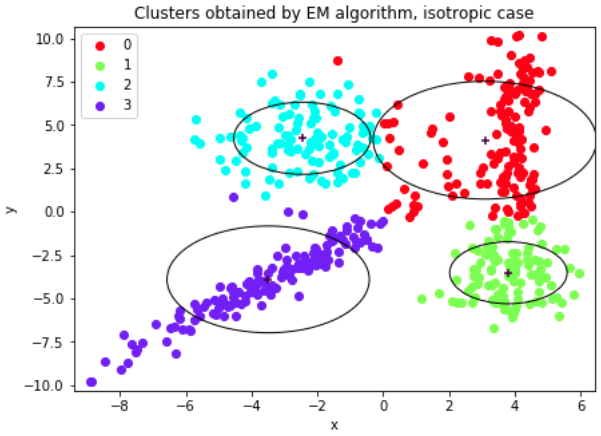
The measures of distortion are:

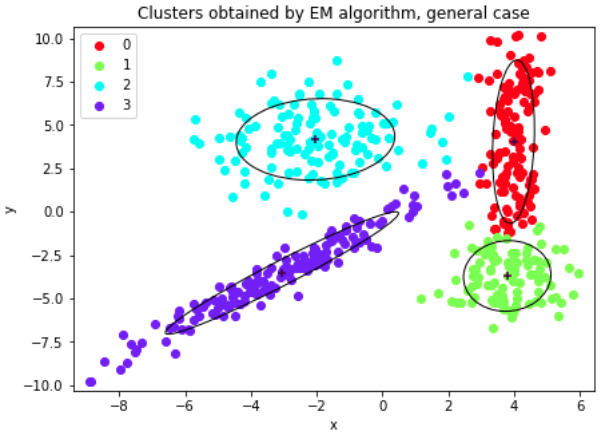
	Train set	Test set
isotropic	3457.9	3354.4
general	4037.4	3642.6

The isotropic model has a lower distortion than the general model. This was to be expected because the general model authorizes more its clusters to have points far from the center.

4 Figures


Clusters obtained by k-Means with k = 4
0
1
2
3
x
y


Clusters obtained by EM algorithm, isotropic case
0
1
2
3
x
y


Clusters obtained by EM algorithm, general case
0
1
2
3
x
y