Master MVA
Probabilistic Graphical Models
HM2 - 21/11/2016
Achari Berrada Youssef

1 Entropy and Mutual Information

1. X a discrete variable on a finite space $\mathcal{X}$ of cardinal k.

   (a) We know that the entropy is
   $$H(X) = -\sum_{x\in\mathcal{X}} p(x)\log(p(x)) \geq 0$$
   Because $0 \leq p(x) \leq 1 \Rightarrow \log(p(x)) \leq 0 \Rightarrow p(x)\log(p(x)) \leq 0$, $\forall x \in \mathcal{X}$.
   So the entropy is greater than or equal to zero.
   If $H(X) = 0$, then $\forall x \in \mathcal{X}, \quad p(x)\log(p(x)) = 0 \Rightarrow \forall x \in \mathcal{X}, p(x) = 1 \,\text{or}\, p(x) = 0$
   As we know that $\sum_{x\in\mathcal{X}} p(x) = 1$, so we conclude that $\exists\, x \in \mathcal{X}$ such that $p(x) = 1$ and null elsewhere.
   So X is constant.

   (b) Let q be the uniform distribution over $\mathcal{X}$. We have
   $$\begin{aligned} D(p||q) &= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{p(x)}{q(x)}\right) \\ &= -H(X) + \log(k) \end{aligned}$$

   (c) So we deduce from a) and b) that :
   $$D(p||q) \leq \log(k)$$

2. (a)
   $$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p(x_1,x_2)\log\left(\frac{p(x_1,x_2)}{p(x_1)p(x_2)}\right) \\ &= \mathbb{E}_{x_1,x_2}\left[-\log\left(\frac{p(x_1)p(x_2)}{p(x_1,x_2)}\right)\right] \\ -\log \text{ is convex} \qquad &\geq -\log\left(\mathbb{E}_{x_1,x_2}\left[\frac{p(x_1)p(x_2)}{p(x_1,x_2)}\right]\right) = -\log(1) = 0 \end{aligned}$$

   (b)
   $$\begin{aligned} I(X_1, X_2) &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p(x_1,x_2)\log\left(\frac{p(x_1,x_2)}{p(x_1)p(x_2)}\right) \\ &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p(x_1,x_2)\log\left(p(x_1,x_2)\right) - \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p(x_1,x_2)\log\left(p(x_1)\right) \\ &\quad - \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p(x_1,x_2)\log\left(p(x_2)\right) \\ &= \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p(x_1,x_2)\log\left(p(x_1,x_2)\right) - \sum_{x_1\in\mathcal{X}_1} p(x_1)\log\left(p(x_1)\right) \\ &\quad - \sum_{x_2\in\mathcal{X}_2} p(x_2)\log\left(p(x_2)\right) \\ I(X_1, X_2) &= H(X_1) + H(X_2) - H(X_1, X_2) \end{aligned}$$

(c) Given X_1, X_2, we have that

$$H(X_1, X_2) \leq H(X_1) + H(X_2)$$

With equality when p_1 and p_2 are independent random variables.
So the joint distribution of the maximal entropy is $p_{1,2} = p_1 p_2$.

2 Conditional independence and factorizations

1. The direct implication is straight forward from the conditional independence axiom :

   If $X \amalg Y|Z$, then $p(x|y,z) = p(x|z)$ for all pairs (y,z) such that $p(y,z) > 0$.

   For the converse, and consider a pair (y,z) such that $p(y,z) > 0$ and assume $p(x|y,z) = p(x|z)$, we have :

   $$p(x,y|z) = p(x|y,z)p(y|z) = p(x|z)p(y|z)$$

   And this implies that $X \amalg Y|Z$.

2. Let's $p \in \mathcal{L}(G)$, we have :

   $$p(x,y,z,t) = p(x)p(y)p(z|x,y)p(t|z)$$

   It is not true that $X \amalg Y|T$. Following a Bayes Ball algorithm, when we shade the node T, a ball can pass from X to through Z then bounce back at Z and finally pass through Y.

3. Statement "$X \amalg Y$ and $X \amalg Y|Z$ then ($X \amalg Z$ or $Y \amalg Z$)"

   (a) Z is a binary variable. Let's note $p(x|z=0) = p_0(x)$, $p(x|z=1) = p_1(x)$, $p(z=1) = q$ and $p(z=0) = 1-q$ with $q \in [0,1]$. We have :

   $$p(x,y) = \sum_z p(x,y,z) = \sum_z p(x,y|z)p(z) = \sum_z p(x|z)p(y|z)p(z) \tag{1}$$

   On the other side,

   $$p(x,y) = p(x)p(y) = \sum_z p(x|z)p(z) \sum_{z'} p(y|z')p(z') \tag{2}$$

   By equaling the equation (1) and (2), we obtain:

   $$((1-q)p_0(x) + qp_1(x)) \,.\, ((1-q)p_0(y) + qp_1(y)) = (1-q)p_0(x)p_0(y) + qp_1(x)p_1(y)$$

   $$\Leftrightarrow \quad p_0(x)p_0(y) + p_0(x)p_1(y) + p_1(x)p_0(y) - p_1(x)p_1(y) = 0$$

   $$\Leftrightarrow \quad (p_0(x) - p_1(x)).(p_0(y) - p_1(y)) = 0$$
   $$\Leftrightarrow \quad p_0(x) = p_1(x) \quad \text{or} \quad p_0(y) = p_1(y)$$

   $$\Leftrightarrow \quad X \amalg Z \quad \text{or} \quad Y \amalg Z$$

   (b) In general, the statement is not true.

3 Distributions factorizing in a graph

1. We know that in a DAG, G = (V,E) , we have that:

$$p(x) = \prod_{k \in V} p(x_k | x_{\pi_k})$$

We know that $\pi_j = \pi_i \cup \{j\}$, then we compute:

$$\begin{aligned} p(x) &= p(x_i|x_{\pi_i}) \quad p(x_j|x_{\pi_j}) && \prod_{k \in V \setminus \{i,j\}} p(x_k|x_{\pi_k}) \\ &= p(x_i|x_{\pi_i}) \quad p(x_j|x_i, x_{\pi_i}) && \prod_{k \in V \setminus \{i,j\}} p(x_k|x_{\pi_k}) \\ &= p(x_j, x_i|x_{\pi_i}) && \prod_{k \in V \setminus \{i,j\}} p(x_k|x_{\pi_k}) \\ &= p(x_j|x_{\pi_i}) \quad p(x_i|x_j, x_{\pi_i}) && \prod_{k \in V \setminus \{i,j\}} p(x_k|x_{\pi_k}) \end{aligned}$$

Which prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

2. G is a directed tree and does not contain any v-structure, hence
$\exists! y \in V, \text{s.t.} |\pi_y| = 0$ and $\forall x \in V \setminus \{y\}, |\pi_x| = 1$.
y is the root of the directed tree.
G' is an undirected tree hence it has only cliques of size 2, because the chromatic number of an undirected tree is 2.

We know that

$$p(x_V) = p(y) \prod_{x \in v_y} p(x|y) \prod_{(x_1,x_2) \in E, \neq y} p(x_2|x_1)$$

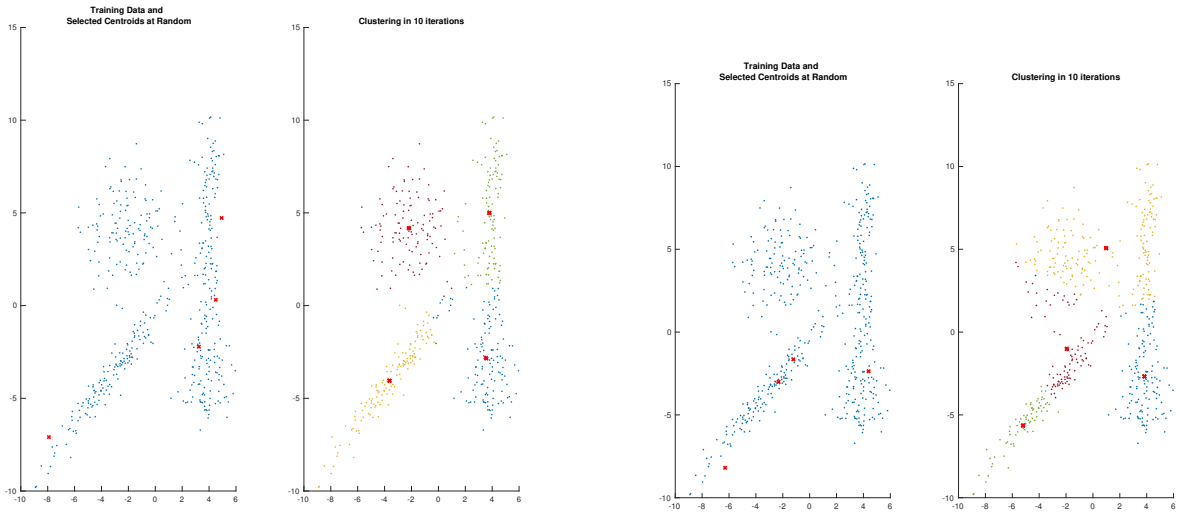
So for all cliques, we define the potential funtion:

- If $C = \{y, x\}$ then $\psi_C(x_C) = p(y)p(x|y)$
- if $C = \{x_1, x_2\}$ such that $x_1 \neq y$ and $x_2 \neq y$ and $(x_1, x_2) \in E$, we define $\psi_C(x_C) = p(x_2|x_1)$.

Then we have proven that $\mathcal{L}(G) = \mathcal{L}(G')$.

4 Implementation - Gaussian mixtures

(a) kMeans is sensible to the randomized centroids initialization, see figure1.



(a) Good kMeans (b) Bad kMeans

Figure 1: kMeans' sensibility on the centroids initialization

(b) The parameters of the EM algorithm are:

$$\theta = (\pi_k, (\mu_k, \Sigma_k))_{k=1,..,K}$$

With $p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$ and $p(x|z;(\mu_k,\Sigma_k)_k) = \sum_k \pi_k \mathcal{N}(x;\mu_k,\Sigma_k)$

We define

$$q(z) = p(z|x,\theta) \qquad \tilde{l}(\theta) = \mathbb{E}_q[\log p(x,z,\theta)]$$

The goal is to maximize $\mathbb{E}_q[\log p(x,z,\theta)]$.
First we initialize $\theta = \theta_0$, and to do so, we use the k-Means result to estimate the parameters of the gaussians. At iteration t, we have :

$$q_{ik}^{(t)} = \mathbb{P}(z_k^{(i)} = 1) = \mathbb{E}_{q_i^{(t)}}\left[z_k^{(i)}\right]$$

1- Expectation step :

$$q_{i,k}^{(t)} \leftarrow \frac{\pi_k^{t-1}\mathcal{N}(x^{(i)}, \mu_k^{(t-1)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{t-1}\mathcal{N}(x^{(i)}, \mu_j^{(t-1)}, \Sigma_j^{(t)})}$$

2- Maximization step :

$$\mu_k^{(t)} = \frac{\sum_i x^{(i)} q_{i,k}^{(t)}}{\sum_i q_{i,k}^{(t)}} \qquad , \qquad \Sigma_k^{(t)} = \frac{\sum_i (x^{(i)} - \mu_k)(x^{(i)} - \mu_k)^T q_{i,k}^{(t)}}{\sum_i q_{i,k}^{(t)}}$$

and

$$\pi_k^{(t)} = \frac{\sum_i x^{(i)} q_{i,k}^{(t)}}{\sum_{i,k'} x^{(i)} q_{i,k'}^{(t)}}$$

See figure 2 for the implementation of this methods in a particular case[1].

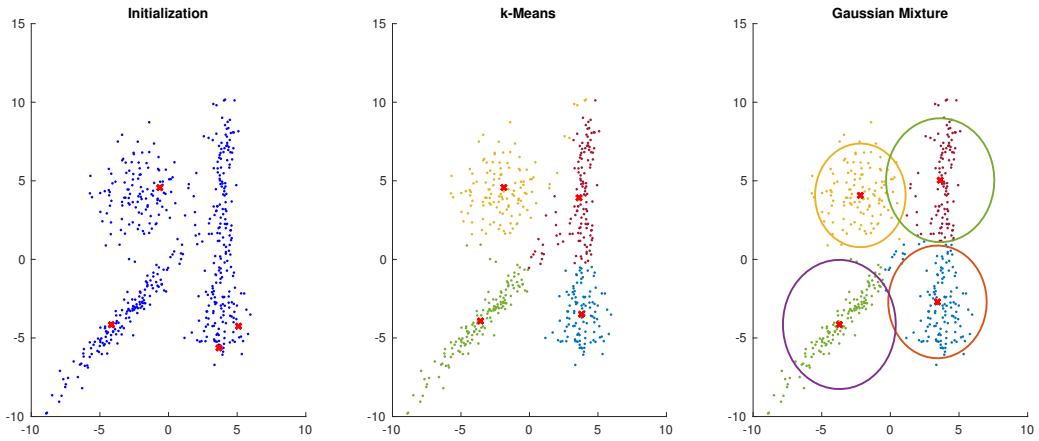


Figure 2: Gaussian Mixture for gaussian with covariance matrices proportional to the identity.

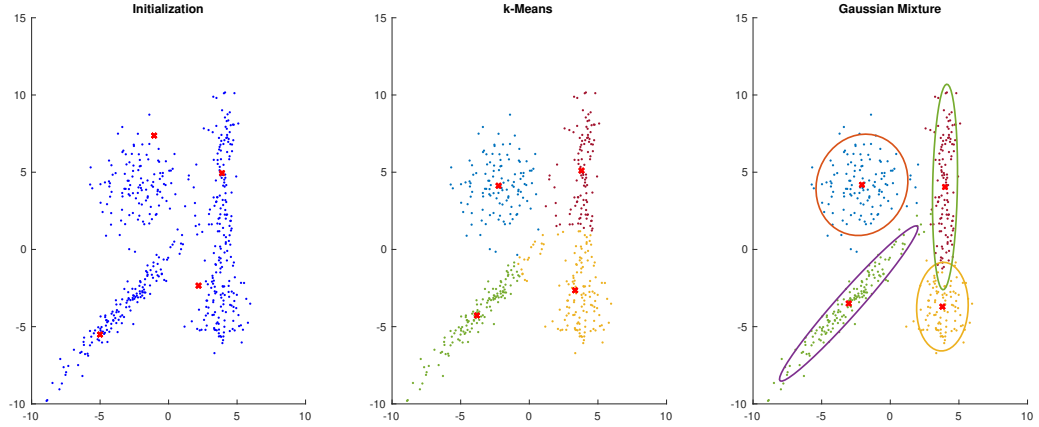


Figure 3: Gaussian Mixture for the general case.

(c) For the general case, see figure 3.

(d) We see clearly from figures 2 and 3, that the gaussian mixture scales better for the general case than for the limited case of identity proportional covariance matrices. This also can be seen very well in the log-likelihood estimations.

Table 1: Log-Likelihood Estimation.

	Particular Case	General Case
Training Data	3.7190	15.6802
Test Data	7.6726	16.3715

[1]I used to plot the Gaussion contour the function PLOT_GAUSSIAN_ELLIPSOID given by Gautam Vallabah in MathWorks.com website.