**Master M2 MVA 2017/2018 - Graphical models**

Homework 2 due November 10th 2017.

SOLUTIONS

# 1 Entropy and Mutual Information

1. (a) We have $H(X) = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$. Since $p(x) \leq 1$, all the terms in the sum are nonpositive, which proves that $H(X) \geq 0$. We have $H(X) = 0$ if and only if, for all $x$, $p(x)\log p(x) = 0$, which entails $p(x) \in \{0, 1\}$ for all $x$. This is possible only if $p$ puts all its mass on a single element $x_0 \in \mathcal{X}$.

   (b),(c) Since $q(x) = \frac{1}{k}$ for all $x \in \mathcal{X}$, $D(p\|q) = -H(X) - \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{k}$. But we have proved in class that $D(p\|q) \geq 0$ with equality if and only if $p = q$ so that $H(X) \leq \log(k)$ with equality if and only if $p$ is the uniform distribution on $\mathcal{X}$.

2. (a) The mutual information is exactly equal to the Kullback-Leibler divergence between $p_{1,2}(\cdot,\cdot)$ and $p_1(\cdot)p_2(\cdot)$ that is $D\big(p_{1,2}(\cdot,\cdot)\|p_1(\cdot)p_2(\cdot)\big)$. It is therefore nonnegative and equal to 0 if and only if $X_1 \perp\!\!\!\perp X_2$.

   (b) We have

$$\begin{aligned}
I(X_1, X_2) &= -H(X_1, X_2) - \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2)(\log p_1(x_1) + \log p_2(x_2)) \\
&= -H(X_1, X_2) - \sum_{x_1\in\mathcal{X}_1} p_1(x_1)\log p_1(x_1) - \sum_{x_2\in\mathcal{X}_2} p_2(x_2)\log p_2(x_2) \\
&= -H(X_1, X_2) + H(X_1) + H(X_2),
\end{aligned}$$

   which proves that $H(X_1, X_2) \leq H(X_1) + H(X_2)$ with equality if and only if $I(X_1, X_2) = 0$, that is when $X_1 \perp\!\!\!\perp X_2$.

   (c) As a consequence the distribution with maximal entropy and given marginal distributions $p_1$ and $p_2$ is the distribution $p_{1,2}(\cdot,\cdot) = p_1(\cdot)p_2(\cdot)$.

# 2 Conditional Independence and factorizations

1. Suppose that $X \perp\!\!\!\perp Y \mid Z$. Then for $(y, z)$ such that $p(x, z) > 0$, we have $p(z) \neq 0$ and $p(y|z) \neq 0$. Then

$$p(x|y,z) = \frac{p(x,y,z)}{p(y,z)} \overset{\text{chain rule}}{=} \frac{p(x,y|z)p(z)}{p(y|z)p(z)} \overset{\substack{\text{factorization form}\\ \text{of cond. ind.}}}{=} \frac{p(x|z)p(y|z)}{p(y|z)} = p(x|z).$$

For the "if" part, let $z$ be such that $p(z) > 0$. If $p(y, z) = 0$, then $p(y|z) = 0$ and $p(x, y, z) = 0$ for any $x$, and thus $p(x, y|z) = 0 = p(x|z)p(y|z)$ trivially. If $p(y, z) > 0$, then we have:

$$p(x, y|z) \overset{\text{chain rule}}{=} p(x|y, z)p(y|z) \overset{\text{assumption}}{=} p(x|z)p(y|z),$$

showing the factorization form of the conditional independence.

2. For $p \in \mathcal{L}(G)$, the factorization is: $p(x, y, z, t) = p(t|z)p(z|x, y)p(x)p(y)$. The answer is no, $X$ and $Y$ have in general no reason to be independent given $T$: take $X$ and $Y$ i.i.d., $Z = 1$ if $X < Y$, $Z = 0$ else, and set $T = Z$. Then clearly $X$ and $Y$ are dependent given $T$. Now, even if $T$ is not deterministic given $Z$ the same problem persists: as a concrete example consider the case of binary variables with $Z = 1$ if and only if $X = Y$ and $p(Z = 1|T = t) = \pi(t)$. Then

$$p(x, y|t) = \sum_{z \in \{0,1\}} \frac{p(x, y, z, t)}{p(t)} = \sum_{z \in \{0,1\}} p(x, y|z)p(z|t).$$

We therefore have $\mathbb{P}(X=1, Y=1|T=t) = \mathbb{P}(X=0, Y=0|T=t) = \pi(t)$ and $\mathbb{P}(X= 0, Y= 1|T=t) = \mathbb{P}(X=1, Y=0|T=t) = 1-\pi(t)$. This conditional distribution of $(X, Y)$ can written as a two-by-two table, and conditional independence would mean that this two-by-two table viewed as a matrix is of rank 1, which entails that its determinant is 0. But this is only true if $\pi(t) = 0.5$ which would force $T$ to be independent from $Z$.

3. (a) If $Z$ is binary, the statement is true. Let's prove it. If $Y$ is a constant r.v. (i.e. $\exists y_0$ s.t. $\mathbb{P}(Y = y_0) = 1$), then $Y$ is trivially independent with any r.v. (verify it!), and so $Y \perp\!\!\!\perp Z$. So we now assume that $Y$ takes at least two distinct values with non-zero probability. For any $y$ such that $p(y) \neq 0$, we have

$$\begin{aligned} p(x) \overset{X \perp\!\!\!\perp Y}{=} \frac{p(x, y)}{p(y)} &= \frac{1}{p(y)} \sum_z p(x, y|z)p(z) \\ &\overset{X \perp\!\!\!\perp Y|Z}{=} \frac{1}{p(y)} \sum_z p(x|z)p(y|z)p(z) = \sum_z p(x|z)p(z|y). \end{aligned}$$

Since $Z$ is binary, we thus have for any $j$ such that $\mathbb{P}(Y = j) \neq 0$,

$$\mathbb{P}(X = i)\mathbb{P}(X = i|Z = 1)\mathbb{P}(Z = 1|Y = j) + \mathbb{P}(X = i|Z = 0)\mathbb{P}(Z = 0|Y = j).$$

Let $u^{(k)}$ be the vector such that $u_i^{(k)} = \mathbb{P}(X = i|Z = k)$ and $v^{(k)}$ be the vector such that $v_j^{(k)} = \mathbb{P}(Z = k|Y = j)$ then

$$A = u^{(0)}{v^{(0)}}^\top + u^{(1)}{v^{(1)}}^\top$$

is the matrix such that $A_{ij} = \mathbb{P}(X = i)$. The columns of $A$ are thus all equal, which means that $u^{(0)}v_j^{(0)} + u^{(1)}v_j^{(1)} = u^{(0)}v_{j'}^{(0)} + u^{(1)}v_{j'}^{(1)}$ for any $j, j'$ such that $\mathbb{P}(Y = j) \neq 0$ and $\mathbb{P}(Y = j') \neq 0$. Since we assume that $Y$ must take at least two different values with non-zero probability, we have that

$$u^{(0)}(v_j^{(0)} - v_{j'}^{(0)}) + u^{(1)}(v_j^{(1)} - v_{j'}^{(1)}) = 0,$$

and so either $u^{(0)}$ and $u^{(1)}$ are collinear or we have both $v_j^{(0)} = v_{j'}^{(0)}$ and $v_j^{(1)} = v_{j'}^{(1)}$.

- In the first case $u^{(0)} = \gamma u^{(1)}$, but we must have $\gamma = 1$ because the entries in $u^{(k)}$ must sum to 1 (it is a probability distribution). So $\mathbb{P}(X|Z = 0) = \mathbb{P}(X|Z = 1)$, implying that $X \perp\!\!\!\perp Z$ (fill in the last details!).
- In the second case, $v_j^{(0)} = v_{j'}^{(0)}$ and $v_j^{(1)} = v_{j'}^{(1)}$ for all pairs $(j, j')$ such that $\mathbb{P}(Y = j) \neq 0$ and $\mathbb{P}(Y = j') \neq 0$. But this means that $\mathbb{P}(Z = 1|Y = j)$ and $\mathbb{P}(Z = 0|Y = j)$ do not depend on $j$ for any $j$ (note in particular that if $\mathbb{P}(Y = j) = 0$ we can set $\mathbb{P}(Z = 1|Y = j) = \mathbb{P}(Z = 1)$ and $\mathbb{P}(Z = 0|Y = j) = \mathbb{P}(Z = 0)$ because on an event of probability 0 the conditional probability can be defined arbitrarily), which means that $Y \perp\!\!\!\perp Z$.

(b) The statement is not true in general. Take $(X, Z_1)$ dependent and $(Y, Z_2)$ dependent such that $(X, Z_1) \perp\!\!\!\perp (Y, Z_2)$. Then define $Z = (Z_1, Z_2)$. We clearly have $X \perp\!\!\!\perp Y$. For the conditional independence, note that $p(x, z) = p(x, z_1)p(z_2)$ and that $p(z) = p(z_1)p(z_2)$ so that $p(x|z) = p(x|z_1)$. Symmetrically $p(y|z) = p(y|z_2)$. Thus

$$p(x, y|z) = \frac{p(x, y, z_1, z_2)}{p(z_1, z_2)} = \frac{p(x, z_1)p(y, z_2)}{p(z_1)p(z_2)} = p(x|z_1)p(y|z_2) = p(x|z)p(y|z),$$

so that $X \perp\!\!\!\perp Y \mid Z$, which completes the proof.

Note that a particular instance of the situation above is the case, where $Z_1 = X$ and $Z_2 = Y$, in which case $Z = (X, Y)$, which provides a simple counterexample, because, conditionally on $Z$, then $X$ and $Y$ are determined and thus independent.

# 3 Distributions factorizing in a graph

1. Let $p \in \mathcal{L}(G)$. We thus have $p(x) = \prod_{k=1}^{n} p(x_k \mid x_{\pi_k})$, where $\pi_k$ denotes the parents of $k$ in $G$. Consider any $x_i, x_j, x_{\pi_i}$ such that $p(x_i, x_j, x_{\pi_i}) \neq 0$. Then by the chain rule (valid for any distribution), we have

$$p(x_i \mid x_{\pi_i})p(x_j \mid x_i, x_{\pi_i}) = p(x_i, x_j \mid x_{\pi_i}) = p(x_j \mid x_{\pi_i})p(x_i \mid x_j, x_{\pi_i}). \qquad (1)$$

As $(i, j)$ is a covered edge, we have $\pi_j = \pi_i \cup \{j\}$. Moreover, by definition of $E'$, we have $\pi'_j = \pi_i$ and $\pi'_i = \pi_j \cup \{j\}$ with $\pi'_i$ the parents of $i$ in $G'$. So note that equation (1) can be interpreted as:

$$p(x_i \mid x_{\pi_i})p(x_j \mid x_{\pi_j}) = p(x_j \mid x_{\pi'_j})p(x_i \mid x_{\pi'_i}).$$

As $\pi'_k = \pi_k$ for any $k \neq i, j$, we can simply swap the two terms for $i$ and $j$ in the product factorization of $p$:

$$p(x) = p(x_i \mid x_{\pi_i})p(x_j \mid x_{\pi_j}) \prod_{k \neq i,j} p(x_k \mid x_{\pi_k}) = p(x_j \mid x_{\pi'_j})p(x_i \mid x_{\pi'_i}) \prod_{k \neq i,j} p(x_k \mid x_{\pi'_k}).$$

If $p(x_i, x_j, x_{\pi_i}) = 0$, then both the LHS and RHS above are equal to zero and so are still equal. We thus have $p \in \mathcal{L}(G')$. By symmetry, we can reverse the argument, and thus $\mathcal{L}(G) = \mathcal{L}(G')$.

2. If $p \in \mathcal{L}(G)$, then $p(x) = \prod_{j=1}^n p(x_j|x_{\pi_j})$ where $|\pi_j| \leq 1$ as $G$ is a directed tree (has no v-structure). Thus denoting $\psi_j(x_j, x_{\pi_j}) = p(x_j|x_{\pi_j})$, $p$ may be written as the Gibbs model $p(x) = \prod_{j=1}^n \psi_j(x_j, x_{\pi_j})$ and thus $p \in \mathcal{L}(G')$.

   For the other direction, we show the result by induction on the size of undirected trees. That is, our induction hypothesis is that for any undirected tree $G' = (V, E')$ with $|V| \leq n$, then $p \in \mathcal{L}(G') \implies p \in \mathcal{L}(G)$ for any directed tree $G$ which is an orientation of $G'$.

   The case $n = 1$ is trivial ($\mathcal{L}(G') =$ all distributions on one node $= \mathcal{L}(G)$).

   So now consider an undirected tree $G' = (V, E')$ with $n > 1$ nodes, and $G = (V, E)$ some directed tree version of $G'$. Let's index the nodes of $V$ from 1 to $n$ so that node $n$ is a leaf which is not the root of the directed tree $G$ and its unique parent is the node $n-1$. For $n > 1$, there exists such a leaf distinct from the root, and for this leaf, we have $(n-1, n) \in E$. Let $p \in \mathcal{L}(G')$, and so we have $p(x) = \frac{1}{Z}\prod_{\{i,j\} \in E'} \psi_{ij}(x_i, x_j)$.

   Let $\tilde{p}$ be the marginal of $p$ on $x_{1:(n-1)}$. Then we have:

$$\tilde{p}(x_{1:(n-1)}) = \frac{1}{Z}\tilde{\psi}(x_{n-1}) \prod_{\{i,j\} \in E' \setminus \{n-1,n\}} \psi_{ij}(x_i, x_j) \quad \text{where} \quad \tilde{\psi}(x_{n-1}) := \sum_{x_n} \psi(x_{n-1}, x_n).$$

   Let $\tilde{G}$ be the subtree of size $n-1$ obtained from $G$ by removing the leaf $n$, and $\tilde{G}'$ its undirected version. From the form above, we see that $\tilde{p} \in \mathcal{L}(\tilde{G}')$. Thus by the induction hypothesis, $\tilde{p} \in \mathcal{L}(\tilde{G})$ and so factorizes as: $\tilde{p}(x_1, \ldots, x_{n-1}) = \prod_{i=1}^{n-1} \tilde{p}(x_i|x_{\pi_i})$. Note that in $G$, $\pi_n = \{n-1\}$; we thus define $f(x_n, x_{\pi_n})$ through

$$f_n(x_n, x_{\pi_n}) := \begin{cases} \psi_{n-1,n}(x_{n-1}, x_n)/\tilde{\psi}(x_{n-1}) & \text{if } \tilde{\psi}(x_{n-1}) \neq 0 \\ 1/K_n & \text{otherwise} \end{cases}$$

with $K_n$ the number of possible values for $X_n$. We then have, valid for all $x$:

$$p(x) = \tilde{p}(x_1, \ldots, x_{n-1}) f_n(x_n, x_{\pi_n}) = f_n(x_n, x_{\pi_n}) \prod_{i=1}^{n-1} \tilde{p}(x_i | x_{\pi_i}).$$

Now since $\sum_{x_n} f_n(x_n, x_{\pi_n}) = 1$, we have that $p$ satisfies the conditions in the definition of $\mathcal{L}(G)$, and thus $p \in \mathcal{L}(G)$, completing the induction step and the proof.

We have just shown that oriented and non-oriented trees are *Markov-equivalent*.

# 4 Mixtures of Gaussians

(a) When initializing the centroids of K-means with $K$ random points from the dataset, we obtain in general different results. Most of them are close to the minimum, but some of them may be quite far (see histogram).

(b) The result is close to K-means since we do not take into accounts correlations between variables. The isotropic covariance matrix estimator is (and following the course notations)

$$\Sigma_i^{(t+1)} = \frac{1}{d} \frac{\sum_n \tau_n^{i(t)} \|x_n - \mu_i^{(t+1)}\|^2}{\sum_n \tau_n^{i(t)}}$$

(NB: don't forget to divide by $d$). The other parameters estimate ($\mu_i^{(t+1)}$ and $\pi_i^{(t+1)}$) during the M-step are the same as seen in class.

A reasonable estimate for the value of the latent variable for each $n$ can be made by maximizing the a posteriori probability $p(z_n|x_n)$, i.e., through $\arg\max_{1 \leq i \leq K} \tau_n^i$.

For a standard multivariate Gaussian, i.e., so that $\mu = 0$ et $\Sigma = I_d$, the disk corresponding to 90% of the mass is centered at zero and has radius $R$ so that $P(r^2 \leqslant R^2) = .9$, $r^2$ being the sum of the $d$ squares of independent standard univariate Gaussians. This is by definition a variable with a $\chi^2$-distribution with $d$ degrees of freedom. In the general case, the ellipse is obtained through an affine transformation (see code).

(c) The covariance matrix estimator is (and following the course notations)

$$\Sigma_i^{(t+1)} = \frac{\sum_n \tau_n^{i(t)} (x_n - \mu_i^{(t+1)})(x_n - \mu_i^{(t+1)})^\top}{\sum_n \tau_n^{i(t)}}$$

(d) We show below the log-likelihood divided by $N_{\text{train}}$ and $N_{\text{test}}$ respectively (we normalize to obtain values which remain small when the number of data points increases and to be able to compare "test" and "train"):

| | Train | Test |
|---|---|---|
| Isotropic | -5.2910 | -5.3882 |
| General | -4.6554 | -4.8180 |

Unnormalized log-likelihoods:

| | Train | Test |
|---|---|---|
| Isotropic | -2.6455 $\times 10^3$ | -2.6941 $\times 10^3$ |
| General | -2.3277 $\times 10^3$ | -2.4090 $\times 10^3$ |

The training log-likelihoods are always greater for more flexible models (the situation may be different for the testing log-likelihoods as the model may be too flexible and we have overfitting). The test log-likelihoods are on average lower than the train ones.