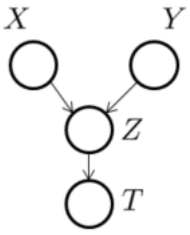
Master M2 MVA 2017/2018 - Graphical models - HWK 2

These exercises are due on November 10th 2017 and should be submitted on the Moodle. They can be done in groups of two students. The write-up can be in French or in English. Please submit your answers as a pdf file that you will name `MVA_DM1_<your_name>.pdf` if you worked alone or `MVA_DM1_<name1>_<name2>.pdf` with both of your names if you worked as a group of two. Indicate your name(s) as well in the documents. Please submit your code as a separate zipped folder and name it `MVA_DM1_<your_name>.zip` if you worked alone or `MVA_DM1_<name1>_<name2>.zip` with both of your names if you worked as group of two. Note that your files should weight no more than 16Mb.

1 Conditional independence and factorizations

1. Prove that $X \perp\!\!\!\perp Y \mid Z$ if and only if $p(x|y,z) = p(x|z)$ for all pairs (y,z) such that $p(y,z) > 0$, using only the three following axioms (a) the definition of conditional independence, (b) the definition of conditional probability via $p(a,b) = p(a|b)p(b)$ and the summation rule $\sum_a p(a|b) = 1$.



2. Consider the directed graphical model G on the right. Write down the implied factorization for any joint distribution $p \in \mathcal{L}(G)$. Is it true that $X \perp\!\!\!\perp Y \mid T$ for any $p \in \mathcal{L}(G)$? Prove or disprove.

3. Let (X, Y, Z) be a r.v. on a finite space. Consider the following statement:

   "If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp Y$ then ($X \perp\!\!\!\perp Z$ or $Y \perp\!\!\!\perp Z$)."

   (a) Is this true if one assumes that Z is a binary variable? Prove or disprove.

   (b) Is the statement true in general[1]? Prove or disprove.

2 Distributions factorizing in a graph

1. Let $G = (V, E)$ be a DAG. We say that an edge $i \to j \in E$ is a *covered edge* if and only if $\pi_j = \pi_i \cup \{i\}$; let $G' = (V, E')$, with $E' = (E \backslash \{i \to j\}) \cup \{j \to i\}$. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

2. Let G be a directed tree and G' its corresponding undirected tree (where the orientation of edges is ignored). Recall that by the definition of a directed tree, G does not contain any v-structure. Prove that $\mathcal{L}(G) = \mathcal{L}(G')$.

[1] This question is harder...

3 Entropy and Mutual Information

1. Let X be a discrete random variable on a finite space $\mathcal{X}$ with $|\mathcal{X}| = k$. Its entropy is defined as

$$H(X) = -\sum_{x\in\mathcal{X}} p_X(x)\log p_X(x), \qquad \text{with} \qquad p_X(x) := \mathbb{P}(X = x),$$

where here log denotes the natural logarithm[2] (to keep notations simple), and where we extend by continuity the function $z \mapsto z\log z$ in 0 so that $0\log 0 = 0$.

Let p and q be two distributions defined on $\mathcal{X}$. The Kullback-Leibler divergence between p and q is denoted $D(p\|q)$ and is defined as follows:

$$D(p\|q) = \begin{cases} +\infty & \text{if } \exists x \in \mathcal{X}, \text{ such that } q(x) = 0 \text{ and } p(x) \neq 0, \\ \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{q(x)} & \text{otherwise,} \end{cases}$$

with the convention that $0\log\frac{0}{0} = 0$. We will prove in class that for all pairs of distributions (p, q), $D(p\|q) \geq 0$ and that $D(p\|q) = 0$ if and only if $p = q$. You may therefore use this property.

   (a) Prove that the entropy $H(X)$ is greater than or equal to zero, with equality if and only if X is a constant with probability 1.

   (b) Denote by p_X the distribution of X and q the uniform distribution on $\mathcal{X}$. What is the relation between the Kullback-Leibler divergence $D(p_X\|q)$ and the entropy $H(X)$ of the distribution p_X?

   (c) Deduce an upper bound on the entropy that depends on k.

2. We consider a pair of discrete random variables (X_1, X_2) defined over the finite set $\mathcal{X}_1 \times \mathcal{X}_2$. Let $p_{1,2}$, p_1 and p_2 denote respectively the joint distribution of (X_1, X_2), the marginal distribution of X_1 and the marginal distribution of X_2. The mutual information $I(X_1, X_2)$ is defined as

$$I(X_1, X_2) = \sum_{(x_1,x_2)\in\mathcal{X}_1\times\mathcal{X}_2} p_{1,2}(x_1, x_2)\log\frac{p_{1,2}(x_1, x_2)}{p_1(x_1)p_2(x_2)}.$$

   (a) Prove that $I(X_1, X_2) \geq 0$.

   (b) Show that $I(X_1, X_2)$ can be expressed as a function of $H(X_1), H(X_2)$ and $H(X_1, X_2)$ where $H(X_1, X_2)$ is the entropy of the random variable $X = (X_1, X_2)$.

   (c) What is the joint distribution $p_{1,2}$ of maximal entropy with given marginals p_1 and p_2?

[2] In the context of coding theory the logarithm used in the definition of the entropy is the logarithm in base 2, and the entropy is measured in *bits*. Here, since we use the natural logarithm, the entropy is measured in *nats*.

4 Implementation - Gaussian mixtures

The file "EMGaussian.data" contains sample of data x_n where $x_n \in \mathbb{R}^2$. The goal of this exercise is to implement the EM algorithm for certain mixtures of K Gaussians in $\mathbb{R}^d$ (here $d = 2$ and $K = 4$), for i.i.d. data. (NB: in this exercise, no need to prove any of the formulas used in the algorithms except for question (b)).

The choice of the programming language is yours (we however recommend Matlab, Scilab, Octave, Python or R). The source code should be handed in along with results. However all the requested figures should be printed on paper or part of a pdf file which is turned in, with clear titles that indicate what the figures represent. The discussions may of course be handwritten.

(a) Implement the K-means algorithm. Represent graphically the training data, the cluster centers, as well as the different clusters. Try several random initializations and compare results (centers and distortion measures).

(b) Consider a Gaussian mixture model in which the covariance matrices are proportional to the identity. Derive the form of the M-step updates for this model and implement the corresponding EM algorithm (using an initialization with K-means).

Represent graphically the training data, the centers, as well as the covariance matrices (an elegant way is to represent the ellipse that contains a certain percentage, e.g., 90%, of the mass of the Gaussian distribution).

Estimate and represent (e.g. with different colors or different symbols) the latent variables for all data points (with the parameters learned by EM).

(c) Implement the EM algorithm for a Gaussian mixture with general covariance matrices. Represent graphically the training data, the centers, as well as the covariance matrices.

Estimate and represent (e.g. with different colors or different symbols) the latent variables for all data points (with the parameters learned by EM).

(d) Comment the different results obtained in earlier questions. In particular, compare the log-likelihoods of the two mixture models on the training data, as well as on test data (in "EMGaussian.test").