Master M2 MVA 2013/2014 - Graphical models

These exercises are due on November 6th 2013.

They can be done in groups of two students.

The writeup may be either in French or in English.

If you plan on sending us your homework as a pdf, please name it

`MVA_DM2_<your name>.pdf` ou `MVA_DM2_<name1>_<name2>.pdf`

1 Distributions factorizing in a graph

(a) Show the following propositions stated in class :

- Let $G = (V, E)$ be a DAG and let $i \to j$ be a *covered edge*, i.e. such that $\pi_j = \pi_i \cup \{i\}$; let $G' = (V, E)$, with $E' = (E \backslash \{i \to j\}) \cup \{j \to i\}$, then we have $\mathcal{L}(G) = \mathcal{L}(G')$.
- Let G be a directed tree without v-structures and G' the undirected tree with the same edges (the symmetrized graph) then $\mathcal{L}(G) = \mathcal{L}(G')$.

(b) Find the smallest undirected graph G such that there exists no directed graph G' such that $\mathcal{L}(G) = \mathcal{L}(G')$.

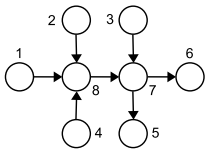
2 d-separation

(a) Let G be a DAG and G_M its moral graph. Let A, B, S be three sets of disjoints vertices. Show that if A and B are separated by S in G_M, they are d-separated by S in G.

(b) Let A, B, S and T non-empty subsets of vertices of a DAG G. If A and B are d-separated by S and A and S are d-separated by T, are A and B d-separated by T ?

(c) Let G be the graph below. Among the following (conditional) independence statements, which are true for all distributions $p \in \mathcal{L}(G)$?

- $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid X_3$
- $X_{\{1,2\}} \perp\!\!\!\perp X_4 \mid X_5$
- $X_1 \perp\!\!\!\perp X_6 \mid X_{\{2,4,7\}}$



3 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).

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) Implement the EM algorithm for a Gaussian mixture with covariance matrices proportional to identity (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 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 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").