Master M2 MVA 2014/2015 - Graphical models
Take Home Exam

Due on Wednesday January 7th, 2015.

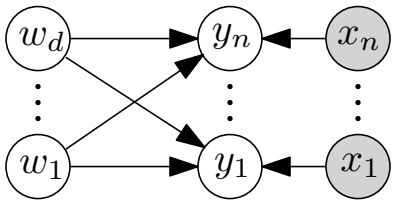
You are requested to work on this take home exam **alone**, without exchanging information with other students.

Please submit this exam as a pdf file on the noodle, please name the file

`MVA_DM3_<your_name>.pdf`

1 Bayesian least-squares regression

We consider n observations $x_i \in \mathbb{R}^d$, $i = 1, \dots, n$, which are assumed to be deterministic, and the following graphical model with the distributions below :



$$\begin{aligned} y_i &\sim \text{Normal}(x_i^\top w, \sigma^2), \quad i \in \{1, \dots, n\} \\ w_j &\sim \text{Normal}(0, \eta_j), \quad j \in \{1, \dots, d\}. \end{aligned}$$

This model can be understood as a Bayesian formulation for least-square regression, where w are the parameters, treated as random variables, $\prod_i p(w_i|\eta_i)$ is the a priori distribution over w, with hyperparameters $(\eta_1, \dots, \eta_d)$. We consider the concatenation of the data in the matrix $X \in \mathbb{R}^{n \times d}$ and the vector $y \in \mathbb{R}^n$. In this exercise, the weight vector $w \in \mathbb{R}^d$ is treated as a latent random variable and we will focus on the dependence on the hyperparameters $\eta \in \mathbb{R}_+^d$ and $\sigma^2 \in \mathbb{R}_+$.

(a) Express the joint distribution of y and w using vector notations (and not as a function of the individual y_i and w_j). All computations in this exercise have to be done with the compact notations X, w and y.

(b) Characterize the marginal distribution $p(y)$ of y and express its parameters as functions of X, σ^2 and η.

(c) Write down the log-likelihood $\log p(y)$.

(d) Characterize the conditional distribution $p(w|y)$ of w given y and express its parameters as functions of X, σ^2 and η.

(e) Give explicit formulas for the EM algorithm to learn parameters $\sigma^2 \in \mathbb{R}^+$ and $\eta \in \mathbb{R}_+^d$.

(f) If some of the parameters η_j are equal to zero, what does the model achieve ?

2 Learning graphical model structures

Throughout this exercise, we consider d discrete random variables $X = (X_1, \dots, X_d)$ taking k values.

(a) For a distribution p on X and given a directed acyclic graph on $\{1, \dots, d\}$, give an expression of the distribution q that factorizes according to G, with minimum KL divergence $D(p||q)$ with p.

(b) Give a simple expression for the optimal KL divergence : $\min_{q \in \mathcal{L}(G)} D(p||q)$ in terms of entropies of subsets $H(X_A)$ for $A \subset \{1, \dots, d\}$, where $H(X_A)$ is the entropy of X_A when X follows the distribution p, that is $H(X_A) = -\sum_{x_A} p(x_A) \log p(x_A)$.

(c) We assume that we are given i.i.d. observations $X^i = (X_1^i, \dots, X_d^i)$ for $i \in \{1, \dots, n\}$. Using (a) and (b), give an expression of the maximum likelihood distribution among all distributions factorizing in a given DAG G.

(d) In the remaining questions, we consider learning the directed graph G defining the structure of the graphical model itself directly from the data. What are the graphs G that maximize the likelihood ? Does this provide reasonable way to learn the graph ?

(e) Compute the number of scalar parameters $c(G)$ which one has to give to specify a probability distribution for the random variable $X = (X_1, \dots, X_d)$ that factorizes according to the directed graph G.

(f) From now on, we consider maximizing with respect to the choice of the directed graph G the sum of the log-likelihood and of a penalty proportional to $c(G)$. Show that two DAGs that only differ by the reversal of single edge that is not involved in any v-structure have the same cost.

(g) BONUS : If the DAG is restricted to be a directed tree or a forest (that is a graph that has at most a parent per node), describe an algorithm that finds the optimal structure.

3 HMM - Implementation

We consider the same training data as in the previous homework, provided as the "EMGaussienne.dat" file (and we will test on the corresponding testing data from the "EMGaussienne.test" file), but this time we use an HMM model to account for the possible temporal structure of the data. The data are of the form $u_t = (x_t, y_t)$ where $u_t = (x_t, y_t) \in \mathbb{R}^2$, for $t = 1, \dots, T$. The goal of this exercise is to implement the probabilistic inference algorithm and the EM algorithm to learn parameters as well as the Viterbi algorithm. It is recommended to make use of the code of the previous homework.

We consider the following HMM model : the chain (q_t) has $K = 4$ possible states, with an initial probability distribution $\pi \in \mathbb{R}^4$ and a probability transition matrix

$A \in \mathbb{R}^{4\times 4}$, and conditionally on the current states we have observations obtained from Gaussian emission probabilities $u_t|q_t = i \sim \mathcal{N}(\mu_i, \Sigma_i)$.

1. Implement the recursions α et β seen in class (and that can be found in the polycopié as well) to compute $p(q_t|u_1, \ldots, u_T)$ and $p(q_t, q_{t+1}|u_1, \ldots, u_T)$.
2. Using the same parameters for the means and covariance matrix of the 4 Gaussians as the ones obtained in the previous homework, taking a uniform initial probability distribution π, and setting A to be the matrix with diagonal coefficients $A_{ii} = \frac{1}{2}$ and off-diagonal coefficients $A_{ij} = \frac{1}{6}$ for all $(i, j) \in \{1, \ldots, 4\}^2$, compute α_t and β_t for all t on the test data ("EMGaussienne.test" file) and compute $p(q_t|u_1, \ldots, u_T)$. Finally, represent $p(q_t|u_1, \ldots, u_T)$ for each of the 4 states as a function of time for the 100 first datapoints in the file. Note that only the 100 first points should be plotted by that filtering should be done with all the data (i.e. $T = 500$). This will be the same for the subsequent questions. (In Matlab the command subplot might be handy to make long horizontal plots.)
3. Derive the estimation equations of the EM algorithm.
4. Implement the EM algorithm to learn the parameters of the model ($\pi, A, \mu_k, \Sigma_k, k = 1 \ldots, 4$). The means and covariances could be initialized with the ones obtained in the previous homework. Learn the model from the training data in"EMGaussienne.dat".
5. Plot the log-likelihood on the train data "EMGaussienne.dat" and on the test data "EMGaussienne.test" as a function of the iterations of the algorithm. Comment.
6. Return in a table the values of the log-likelihoods of the Gaussian mixture models and of the HMM on the train and on the test data. Compare these values. Does it make sense to make this comparison ? Conclude. Compare these log-likelihoods as well with the log-likelihoods obtained for the different models in the previous homework.
7. Implement Viterbi decoding (aka MAP inference) to estimate the most likely sequence of states, i.e. $\arg\max_q p(q_1, \ldots, q_T|y_1, \ldots, y_T)$.

   For the set of parameters learned with the EM algorithm, compute the most likely sequence of states with the Viterbi algorithm and represent the data in 2D with the cluster centers and with markers of different colors for the datapoints belonging to different classes.
8. For the datapoints in the test file "EMGaussienne.test", compute the marginal probability $p(q_t|u_1, \ldots, u_T)$ for each point to be in state $\{1, 2, 3, 4\}$ for the parameters learned on the training set. For each state plot the probability of being in that state as a function of time for the 100 first points (i.e., as a function of the datapoint index in the file).
9. For each of these same 100 points, compute their most likely state according to the marginal probability computed in the previous question. Make a plot

representing the most likely state in $\{1, 2, 3, 4\}$ as function of time for these 100 points.

10. Run Viterbi on the test data. Compare the most likely sequence of states obtained for the 100 first data points with the sequence of states obtained in the previous question. Make a similar plot. Comment.
11. In this problem the number of states was known. How would you choose the number of states if you did not know it ?