Fundamentals of Probabilistic Data Mining

Jean-Baptiste Durand

Duration: 3 hours. Every document allowed. No computer or calculator or mobile phone allowed.

The two exercises are independent. If you cannot prove a statement, you can admit it in the next questions (not in the previous questions!). The approximate weight of each exercise in the final mark is given as a %.

Exercise 1 [70%]: We aim at modelling biological sequences of DNA. The observations $x_1^n = (x_1, \ldots, x_n)$ take values into a finite set {A, C, G, T}. We seek to segment the sequence into homogeneous zones. Biologists emphasize that homogeneous zones are essentially characterized by the way the symbols A, C, G and T success to each other within a given zone. The proportion of these symbols within each zone is not so much relevant. They would like to know whether it makes sense to use a hidden Markov chain model to detect homogeneous zones.

Modelling

1) We recall that the graph in Fig. 1 is a perfect independence map for hidden Markov chains. Explain why this model does not fit the requirements for biologists.

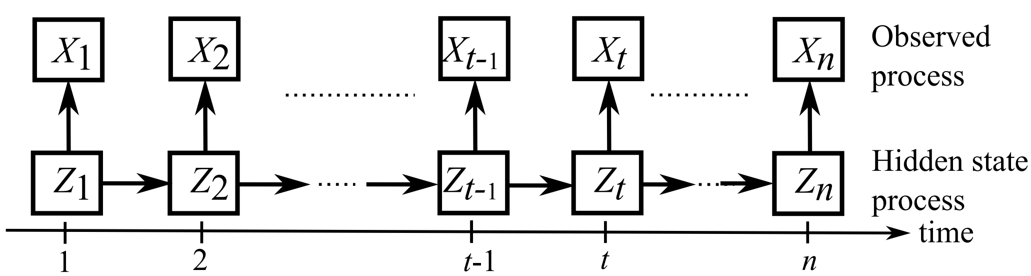

Figure 1: *Independence map for hidden Markov chains*

2) We propose another hidden Markov model that has Fig. 2 as a perfect map. This model is referred to as M1M1. Explain whether the latter corresponds better or not than a hidden Markov chain to the problem raised by biologists.

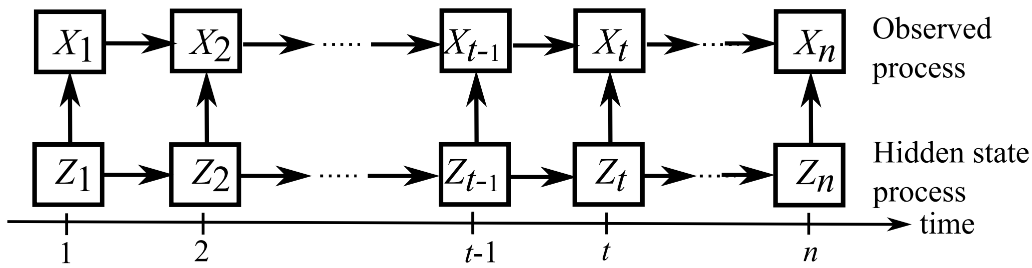

Figure 2: *Perfect independence map for model M1M1*

3) Let $X_1^n = (X_1, \ldots, X_n)$ denote the observed process, $Z_1^n = (Z_1, \ldots, Z_n)$ the hidden state process, $\{1, \ldots, M\}$ the possible values of the observed process and $\{1, \ldots, K\}$ the possible values of the hidden states.

(a) Show that assuming homogeneity of the process, the canonical model parametrization is:

- $(\pi_i)_{i=1,\ldots,K}$ with $\pi_i = P(Z_1 = i)$;
- $(B_{ix})_{\substack{i=1,\ldots,K \\ x=1,\ldots,M}}$ with $B_{ix} = P(X_1 = x | Z_1 = i)$;
- $(P_{ij})_{i,j=1,\ldots,K}$ with $P_{ij} = P(Z_{t+1} = j | Z_t = i)$;
- $(A_{xj,y})_{\substack{x,y=1,\ldots,M \\ j=1,\ldots,K}}$ with $A_{xj,y} = P(X_{t+1} = y | Z_{t+1} = j, X_t = x)$.

Let $\boldsymbol{\lambda}$ denote the set of model parameters.

(b) Show that the completed likelihood writes

$$P_{\boldsymbol{\lambda}}(x_1^n, z_1^n) = \prod_i \pi_i^{\mathbb{I}_{\{z_1=i\}}} \prod_{i,x} B_{ix}^{\mathbb{I}_{\{z_1=i,x_1=x\}}} \prod_{t=1}^{n-1} \prod_{x,j,y} A_{xj,y}^{\mathbb{I}_{\{x_t=x,z_{t+1}=j,x_{t+1}=y\}}} \prod_{t=1}^{n-1} \prod_{i,j} P_{ij}^{\mathbb{I}_{\{z_t=i,z_{t+1}=j\}}}$$

where $\mathbb{I}_{\mathcal{E}}$ refers to the indicator function, $\mathcal{E}$ being some set.

4) Discuss the following statement: "$(X_t)_{t\in\mathbb{N}}$ is some non-homogeneous Markov chain".

EM algorithm: M step

We consider maximum likelihood estimation of parameter $\boldsymbol{\lambda}$ with the EM algorithm, using observed sequence x_1^n. Let $\boldsymbol{\lambda}^{(m)}$ denote the parameter value at iteration m of the algorithm.

Let $\gamma_t^{(m)}(i)$ denote $P_{\boldsymbol{\lambda}^{(m)}}(Z_t = i | X_1^n = x_1^n)$ and $\xi_t^{(m)}(i,j)$ denote $P_{\boldsymbol{\lambda}^{(m)}}(Z_t = i, Z_{t+1} = j | X_1^n = x_1^n)$.

5) Show that iteration m of the EM algorithm resorts to maximizing the following function with respect

to $\boldsymbol{\lambda} = (\pi, B, P, A)$

$$Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(m)}) = \sum_i \gamma_1^{(m)}(i) \ln(\pi_i) + \sum_{i,x} \mathbb{1}_{\{x_1 = x\}} \gamma_1^{(m)}(i) \ln(B_{ix})$$
$$+ \sum_{t=1}^{n-1} \sum_{x,j,y} \mathbb{1}_{\{x_t = x, x_{t+1} = y\}} \gamma_{t+1}^{(m)}(j) \ln(A_{xj,y}) + \sum_{t=1}^{n-1} \sum_{i,j} \xi_t^{(m)}(i,j) \ln(P_{ij}).$$

6) We assume (since this a result from the course) that the solution of the maximisation problem

$$\arg \max_{\substack{(p_1,\ldots,p_K) \in \mathbb{R}_+^K \\ \sum_{k=1}^{K} p_k = 1}} \sum_{t=1}^{n} \sum_{k=1}^{K} \eta_{t,k} \ln(p_k) \text{ is } \hat{p}_k = \frac{\sum_{t=1}^{n} \eta_{t,k}}{\sum_{\ell=1}^{K} \sum_{t=1}^{n} \eta_{t,\ell}}.$$

Provide the re-estimation formulas for P and A at iteration m of the EM algorithm. Justify your answer but do not compute gradients nor partial derivatives of $Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}^{(m)})$.

EM algorithm: E step

For the sake of simplicity, let denote $P = P_{\boldsymbol{\lambda}^{(m)}}$.

We are trying to develop a *forward* recursion to compute $\alpha_t(i) = P(Z_t = i, X_1^t = x_1^t)$.

7) For every i, give an expression of $\alpha_1(i)$ as a function of the model parameters and data.

8) Draw the minimal undirected graph required to decide whether $Z_{t+1} \perp\!\!\!\perp X_1, \ldots, X_t | Z_t$ and provide some justification for the graph you propose. Deduce some expression of $P(Z_{t+1} = j | Z_t = i, X_1^t = x_1^t)$ as a function of model parameters and data.

9) Similarly, draw the minimal undirected graph required to decide whether $X_{t+1} \perp\!\!\!\perp X_1, \ldots, X_{t-1} | X_t, Z_{t+1}$ and provide some justification for the graph you propose.

10) Deduce from the last two questions that

$$\forall j, \, \forall t < n, \quad \alpha_{t+1}(j) = \sum_i A_{x_t j, x_{t+1}} \, P_{ij} \, \alpha_t(i).$$

We assume that similarly, some *backward* recursion allows us to compute
$\beta_t(i) = P(X_{t+1}^n = x_{t+1}^n | X_t = x_t, Z_t = i)$ as:

$$\beta_{t-1}(h) = \sum_i \beta_t(i) \, A_{x_{t-1} i, x_t} \, P_{hi}.$$

11) Give an algorithm with polynomial time complexity for computing the likelihood of some given parameter $\boldsymbol{\lambda}$ on x_1^n (the complexity being a function of n and K). Provide an asymptotic equivalent of that complexity.

12) Show that with the data, parameter $\boldsymbol{\lambda}^{(m)}$ and the outputs of the *forward* and *backward* recursions, all quantities required to implement the M step of the EM algorithm can be computed with polynomial time complexity. In particular, provide some formulas to compute $\xi_t(i,j)$ and $\gamma_t(i)$.

Exercise 2 [30%]: probabilistic graphical models

Multivariate Gaussians

1) We consider a Gaussian vector $(X_1, \ldots, X_6)$ with 0 mean and covariance matrix Σ defined as

$$\Sigma = \begin{pmatrix} 1 & 0 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0 & 1 & 0.1 & 0.1 & 0.1 & 0.1 \\ 0.1 & 0.1 & & & & \\ 0.1 & 0.1 & & & & \\ 0.1 & 0.1 & & A & & \\ 0.1 & 0.1 & & & & \end{pmatrix} \quad \text{with } A = \begin{pmatrix} a+e & -\frac{1}{3}+e & -\frac{1}{3}+e & a-1+e \\ \frac{1}{3}+e & a+e & a-1+e & -\frac{1}{3}+e \\ \frac{1}{3}+e & a-1+e & a+e & -\frac{1}{3}+e \\ a-1+e & -\frac{1}{3}+e & -\frac{1}{3}+e & a+e \end{pmatrix}, \; a = \frac{7}{6} \text{ and } e = \frac{2}{100}.$$

We give the upper triangular part of the symmetric matrix Σ^{-1}:

$$\begin{pmatrix} 1.06 & 0.06 & -0.15 & -0.15 & -0.15 & -0.15 \\ & 1.06 & -0.15 & -0.15 & -0.15 & -0.15 \\ & & 1 & \frac{1}{4} & \frac{1}{4} & 0 \\ & & & 1 & 0 & \frac{1}{4} \\ & & & & 1 & \frac{1}{4} \\ & & & & & 1 \end{pmatrix} \quad \text{and } (A - e\,\mathbb{1}_4)^{-1} = \begin{pmatrix} 1 & \frac{1}{4} & \frac{1}{4} & 0 \\ \frac{1}{4} & 1 & 0 & \frac{1}{4} \\ \frac{1}{4} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{4} & \frac{1}{4} & 1 \end{pmatrix}$$

where $\mathbb{1}_4$ is the 4 by 4 matrix full of ones.

Using the formulas of conditioned Gaussian vectors,

$$\mu_{A|B} = \mu_A - \Sigma_{A,B}\Sigma_B^{-1}(x_b - \mu_B)$$
$$\Sigma_{A|B} = \Sigma_A - \Sigma_{A,B}\Sigma_B^{-1}\Sigma_{B,A},$$

find a perfect undirected I-MAP for $P(X_3, \ldots, X_6 | X_1, X_2)$.

2) Draw some minimal undirected I-MAP for $P(X_1, \ldots, X_6)$. Is it a perfect map? Provide some detailed justification for your answers (several lines of comments required).

3) Draw some minimal directed I-MAP for $P(X_1, \ldots, X_6)$. Is it a perfect map? Provide some detailed justification for your answers (several lines of comments required).

I-equivalence PDAGs

4) We now consider distributions having DAG $\mathcal{G}_1$ or PDAG $\mathcal{G}_2$ as a perfect map (Fig. 3). Draw the I-equivalence PDAG for $\mathcal{G}_1$ (justify your answer).

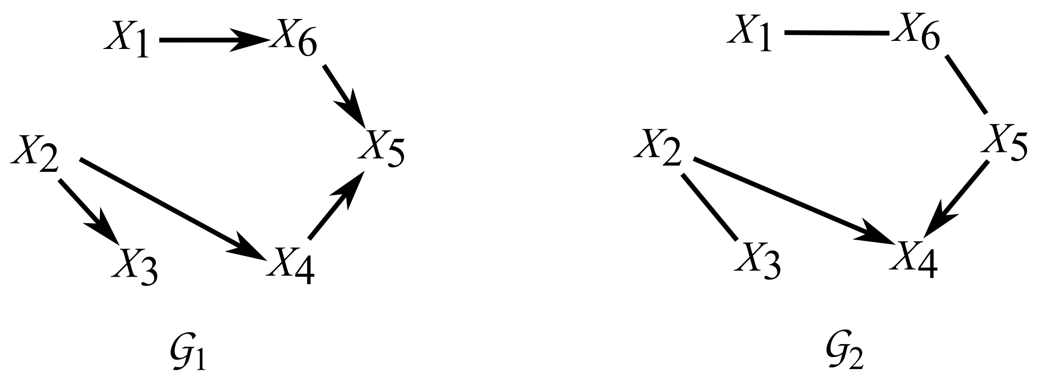

Figure 3: *Perfect independence maps* $\mathcal{G}_1$ *and* $\mathcal{G}_2$*.*

5) Draw an array containing in the first line every possible marginal or conditional independence relationship that does not hold in both $\mathcal{G}_1$ and $\mathcal{G}_2$. For each of them, put True in the second line iff the relationship holds in $\mathcal{G}_1$ (False otherwise). Put True in the third line iff the relationship holds in $\mathcal{G}_2$ (False otherwise).