Probabilistic Graphical Models: Homework 3

Petrovich Mathis, Bricout Raphaël

Implementation - HMM

Exercice 1.2

Please find all details in annex.
We use α and β to compute $p(q_t \,|\, u_1, ...u_T)$ and $p(q_t, q_{t+1} \,|\, u_1, ...u_T)$, and with these quantities we can compute the complete log-likelihood, for $u = (u_1, ...u_T)$:

$$l_c(\theta) = \log\left(p(q_1)\prod_{t=1}^{T-1} p(q_{t+1} \mid q_t)\prod_{t=1}^{T} p(u_t \mid q_t)\right)$$

$$= \sum_{k=1}^{K} \delta(q_1 = k)\log(\pi_k) + \sum_{t=1}^{T-1}\sum_{i,j=1}^{K} \delta(q_{t+1} = i, q_t = j)\log A_{i,j} + \sum_{t=1}^{T}\sum_{k=1}^{K} \delta(q_t = k)\log\mathcal{N}(\mu_k, \Sigma_k, u_t)$$

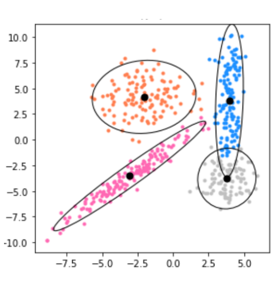
$$\mathbb{E}_q(l_c(\theta)) = \sum_{k=1}^{K} p(q_1 = k \mid u;\theta)\log(\pi_k) + \sum_{t=1}^{T-1}\sum_{i,j=1}^{K} p(q_{t+1} = i, q_t = j \mid u;\theta)\log A_{i,j} + \sum_{t=1}^{T}\sum_{k=1}^{K} p(q_t = k \mid u;\theta)\log\mathcal{N}(\mu_k, \Sigma_k, u_t)$$

With $\log\mathcal{N}(\mu_k, \Sigma_k, u_t) = C + \frac{1}{2}\log\det\Sigma_k^{-1} - \frac{1}{2}(u_t - \mu_k)^T\Sigma_k^{-1}(u_t - \mu_k)$
By Jensen's inequality, $\log p(u) \geq \mathbb{E}_q(l_c(\theta)$, so to maximize the log-likelihood of the data, we will maximize the lower bound $\mathbb{E}_q(l_c(\theta)$ (E step). We can do that with the Lagrangian to update the parameters $\{\pi, A, \Sigma, \mu\}$ (M step) and we get (after some calculations):

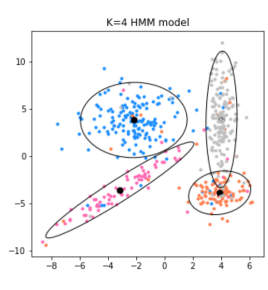
$\pi_k = \frac{p(q_1=k \mid u;\theta)}{\sum_{j=1}^{K} p(q_1=j \mid u;\theta)}$	$A_{i,j} = \frac{\sum_{t=1}^{T-1} p(q_{t+1}=i, q_t=j \mid u;\theta)}{\sum_{l=1}^{K}\sum_{t=1}^{T-1} p(q_{t+1}=l, q_t=j \mid u;\theta)}$
$\mu_k = \frac{\sum_{t=1}^{T} p(q_t=k \mid u;\theta)u_t}{\sum_{t=1}^{T} p(q_t=k \mid u;\theta)}$	$\Sigma_k = \frac{\sum_{i=1}^{T} p(q_t=k \mid u;\theta)(u_t-\mu_k)(u_t-\mu_k)^T}{\sum_{i=1}^{T} p(q_t=k \mid u;\theta)}$

Exercice 1.4

EM



HMM



Exercice 1.5

Log-likelihood		Old EM	HMM
	Train	-3.11	-3.882
	Test	-3.18	-3.90

The local log-likelihood is a bit worse compared to basic EM, it might be due to several reasons:

- a wrong initialization for the max algorithm
- the parameters π, A are not accurate enough
- The model are different.

Annex

Exercise 1.1

Define α and β. With $q_1, ...q_T$ the latent variables that define a Markov Chain and $u_1, ...u_T$ the observed variables (with empirical distributions $u_1, ...u_T$) that can take K values, and with $a \in \mathbb{R}^{4\times 4}$ the transition matrix, we have :

- the forward messages are defined as follow:

$$\mu_{u_0 \to q_0}(q_0) = p(u_0 \mid q_0)$$

$$\mu_{q_0 \to q_1}(q_1) = \sum_{q_0} p(q_1 \mid q_0)\mu_{u_0 \to q_0}(q_0)$$

  And for any t:

$$\mu_{u_t \to q_t}(q_t) = p(u_t \mid q_t)$$

$$\mu_{q_{t-1} \to q_t}(q_t) = \sum_{q_{t-1}} p(q_t \mid q_{t-1})\mu_{q_{t-2} \to q_{t-1}}(q_{t-1})p(q_{t-1} \mid u_{t-1})$$

  And thus α is defined with and $\alpha_0(q_0) = p(q_0)$ and:

$$\alpha_t(q_t) = \mu_{u_t \to q_t}(q_t)\mu_{q_{t-1} \to q_t}(q_t)$$

  Then we have the following recursion:

$$\alpha_{t+1}(q_{t+1}) = p(u_{t+1} \mid q_{t+1}) \sum_{q_t} p(q_{t+1} \mid q_t)\alpha_t(q_t)$$

- we also define the backward message:

$$\beta_t(q_t) = \mu_{q_{t+1} \to q_t}(q_t)$$

  And by recursion, $\beta_T(q_T) = 1$ and:

$$\beta_t(q_t) = \sum_{q_{t+1}} p(q_{t+1} \mid q_t)p(u_{t+1} \mid q_{t+1})\beta_{t+1}(q_{t+1})$$

With these two recursive formulas we can compute α and β.
Once we have this, we can easily estimate $p(q_t \mid u_1, ...u_T)$ and $p(q_t, q_{t+1} \mid u_1, ...u_T)$ with:

$$p(q_t \mid u_1, ...u_T) = \frac{\alpha_t(q_t)\beta_t(q_t)}{\sum_{q_t} \alpha_t(q_t)\beta_t(q_t)}$$

And:

$$p(q_t, q_{t+1} \mid u_1, ...u_T) = \frac{\alpha_t(q_t)\beta_t(q_t)}{p(u_1, ...u_T)} p(q_{t+1} \mid q_t)p(u_{t+1} \mid q_{t+1})$$

Exercise 1.2

Compute estimation equation of EM.
E-Step : We first compute the complete log-likelihood of the problem:

$$
\begin{aligned}
l_c(\theta) &= \log\left(p(q_1)\prod_{t=1}^{T-1} p(q_{t+1} \mid q_t)\prod_{t=1}^{T} p(u_t \mid q_t)\right) \\
&= \sum_{k=1}^{K} \delta(q_1 = k)\log(\pi_k) + \sum_{t=1}^{T-1}\sum_{i,j=1}^{K} \delta(q_{t+1} = i, q_t = j)\log A_{i,j} + \sum_{t=1}^{T}\sum_{k=1}^{K} \delta(q_t = k)\log \mathcal{N}(\mu_k, \Sigma_k, u_t) \\
\mathbb{E}_q(l_c(\theta)) &= \sum_{k=1}^{K} p(q_1 = k \mid u;\theta)\log(\pi_k) + \sum_{t=1}^{T-1}\sum_{i,j=1}^{K} p(q_{t+1} = i, q_t = j \mid u;\theta)\log A_{i,j} + \sum_{t=1}^{T}\sum_{k=1}^{K} p(q_t = k \mid u;\theta)\log \mathcal{N}(\mu_k, \Sigma_k, u_t)
\end{aligned}
$$

With $u = (u_1, ...u_T)$ and:

$$
\begin{aligned}
\log \mathcal{N}(\mu_k, \Sigma_k, u_t) &= \log \frac{1}{\sqrt{(2\pi)^d|\Sigma_k|}} \exp\left(-\frac{1}{2}(u_t - \mu_k)^T\Sigma_k^{-1}(u_t - \mu_k)\right) \\
&= C + \frac{1}{2}\log\det \Sigma_k^{-1} - \frac{1}{2}(u_t - \mu_k)^T\Sigma_k^{-1}(u_t - \mu_k)
\end{aligned}
$$

For M-Step, we need to maximize this quantity with respect to $\theta = \{\pi, A, \mu, \Sigma\}$ and with some constraint.
Here is the Lagrangian: $\mathcal{L}(\theta, \lambda, h) = l_c(\theta) + \lambda(1 - \sum_{i=1}^{K}\pi_k) + \sum_{j=1}^{K} h_j(1 - \sum_{i=1}^{K} A_{i,j})$
$\mathcal{L}(\theta, \lambda, h)$ is concave in π, λ, h, A, μ and Σ, so we can take gradients to maximize $\mathcal{L}$.
$\nabla_\lambda \mathcal{L}(\theta, \lambda, h) = 0 \Rightarrow \sum_{i=1}^{K}\pi_k = 1$
$\nabla_{\pi_k}\mathcal{L}(\theta, \lambda, h) = \frac{p(q_1=k \mid u;\theta)}{\pi_k} - \lambda = 0 \Rightarrow \pi_k = \frac{p(q_1=k \mid u;\theta)}{\lambda}$
$\sum_{i=1}^{K}\pi_k = 1 \Rightarrow \lambda = \sum_{i=1}^{K} p(q_1 = k \mid u;\theta)$ so

$$\boxed{\pi_k = \frac{p(q_1=k \mid u;\theta)}{\sum_{j=1}^{K} p(q_1=j \mid u;\theta)}}$$

$\nabla_{A_{i,j}}\mathcal{L}(\theta, \lambda, h) = \frac{\sum_{t=1}^{T-1} p(q_{t+1}=i, q_t=j \mid u;\theta)}{A_{i,j}} - h_j = 0 \Rightarrow A_{i,j} = \frac{\sum_{t=1}^{T-1} p(q_{t+1}=i, q_t=j \mid u;\theta)}{h_j}$
$\nabla_{h_j}\mathcal{L}(\theta, \lambda, h) = 0 \Rightarrow \sum_{l=1}^{K} A_{l,j} = 1$ so $h_j = \sum_{l=1}^{K}\sum_{t=1}^{T-1} p(q_{t+1} = l, q_t = j \mid u;\theta)$ so

$$\boxed{\mathrm{A}_{i,j} = \frac{\sum_{t=1}^{T-1} p(q_{t+1}=i, q_t=j \mid u;\theta)}{\sum_{l=1}^{K}\sum_{t=1}^{T-1} p(q_{t+1}=l, q_t=j \mid u;\theta)}}$$

$\nabla_{\mu_k}\mathcal{L}(\theta, \lambda, h) = 0 \Rightarrow D\sum_{t=1}^{T} p(q_t = k \mid u;\theta)\Sigma^{-1}(u_t - \mu_k) = 0$ so
$\sum_{t=1}^{T} p(q_t = k \mid u;\theta)u_t = \mu_k \sum_{t=1}^{T} p(q_t = k \mid u;\theta)$ and finally

$$\boxed{\mu_k = \frac{\sum_{t=1}^{T} p(q_t=k \mid u;\theta)u_t}{\sum_{t=1}^{T} p(q_t=k \mid u;\theta)}}$$

Let's $\Lambda_k = \Sigma_k^{-1}$, we have $\log \mathcal{N}(\mu_k, \Sigma_k, u_t) = C + \frac{1}{2}\log\det \Lambda_k - \frac{1}{2}(u_t - \mu_k)^T\Lambda_k^{-1}(u_t - \mu_k)$
The $\log\det$ function is concave so $\log \mathcal{N}$ is concave. We can maximize it with a gradient.
$\nabla_{\Lambda_k}\mathcal{L}(\theta, \lambda, h) = \frac{1}{2}\sum_{i=1}^{T} p(q_t = k \mid u;\theta)(\Lambda_k^{-1} - (u_t - \mu_k)(u_t - \mu_k)^T) = 0 \Rightarrow$

$$\boxed{\Sigma_k = \frac{\sum_{i=1}^{T} p(q_t=k \mid u;\theta)(u_t-\mu_k)(u_t-\mu_k)^T}{\sum_{i=1}^{T} p(q_t=k \mid u;\theta)}}$$