Probabilistic Graphical Models - Homework 3

Jean-Christophe Corvisier, Amine Kheldouni

5^{th} December 2018

Implementation - Hidden Markov Model

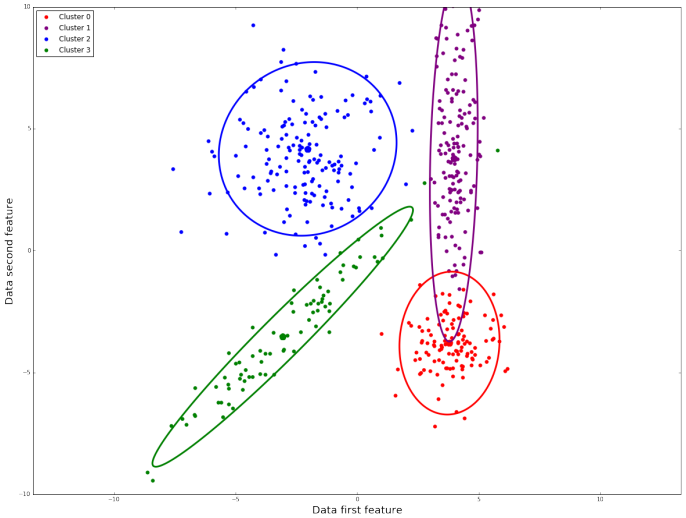
We consider the parameters of our EM algorithm following the handout's notations $\theta = \{\pi_0, a, (\mu_i)_{i\in\{1..4\}}, (\Sigma_i)_{i\in\{1..4\}}\}$. We write the complete log-likelihood $l_c(\theta)$ and we lower bound the log-likelihood by its expectation by Jensen's inequality (as suggested by the lecture).

$$\begin{aligned}
\mathbb{E}\left[l_c(\theta)\right] &= \mathbb{E}\left[\log\left(p(q_0))\prod_{t=0}^{T-1} p(q_{t+1}|q_t)\prod_{t=0}^{T} p(u_t|q_t)\right)\right] \\
&= \sum_{i=1}^{K} p(q_0 = i|u;\theta)\log((\pi_0)_i) + \sum_{t=0}^{T-1}\sum_{i,j=1}^{K} p(q_{t+1}=i, q_t=j|y;\theta)\log(a_{i,j}) \\
&\quad + \sum_{t=0}^{T}\sum_{i=1}^{K} p(q_t = i|y;\theta)\log\left(\frac{1}{(2\pi)|\Sigma_i|^{\frac{1}{2}}}\exp\left(-(u_t-\mu_i)\Sigma_i^{-1}(u_t-\mu_i)\right)\right)
\end{aligned}$$

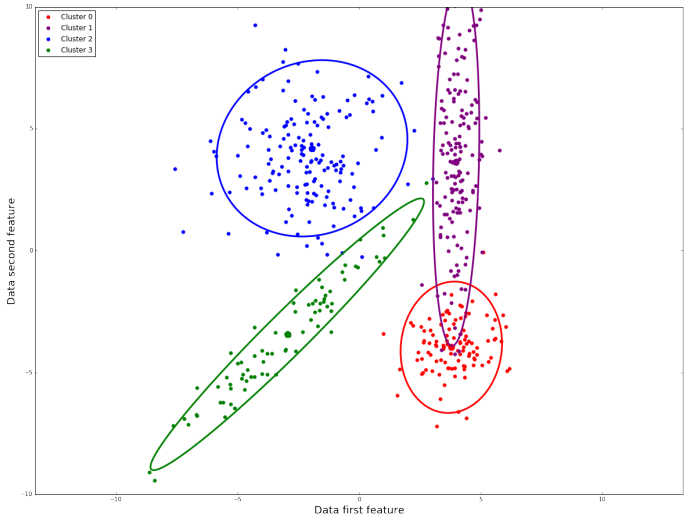
After writing the Lagrangian, we find the following formulas for the M-step at the k^{th} iteration (cf. Appendix) :

$$\pi_{0i}^k = p(z_0 = i|u,\theta^{k-1}) \qquad a_{i,j}^k = \frac{\sum_{t=0}^{T-1} p(z_{t+1}=i, z_t=j|u,\theta^{k-1})}{\sum_{t=0}^{T-1} p(z_{t+1}=i|u,\theta^{k-1})}$$

$$\mu_i^k = \frac{\sum_{t=0}^{T-1} p(z_t=i|u,\theta^{k-1})u_t}{\sum_{t=0}^{T-1} p(z_t=i|u,\theta^{k-1})} \qquad \Sigma_i^k = \frac{\sum_{t=0}^{T-1} p(z_t=i|u,\theta^{k-1})(u_t-\mu_i)(u_t-\mu_i)^T}{\sum_{t=0}^{T-1} p(z_t=i|u,\theta^{k-1})}$$



GMM clustering using EM algorithm (test, $\log\mathcal{L} = -4.91$)



HMM clustering using EM and inference algorithm (test, $\log\mathcal{L} = -3.92$)

We notice from the above figures that the two models tend to have very similar clustering results in our datasets. Both models successfully manage to separate the data into K labels and estimate the parameters (either for Gaussian model or for Hidden Markov model).

Moreover, we notice that the likelihoods on the training data are again higher than those of the test data because the parameters optimization is run according to the training sets. More particularly, the likelihood of HMM is the highest which makes it the better model in this example.

Appendix

Question 2

Demonstration of the formulas for π_0. At step-k, we want to optimize the following quantity w.r.t $\theta = \{\pi_0, a, (\mu_i)_{i\in\{1..4\}}, (\Sigma_i)_{i\in\{1..4\}}\}$:

$$\mathbb{E}_q[l_c(\theta^{k-1})] = \sum_{i=1}^{K} p(q_0 = i|u;\theta^{k-1}) \log((\pi_0)_i) + \sum_{t=0}^{T-1}\sum_{i,j=1}^{K} p(q_{t+1} = i, q_t = j|y;\theta^{k-1}) \log(a_{i,j}) + \sum_{t=0}^{T}\sum_{i=1}^{K} p(q_t = i|y;\theta^{k-1})$$

The variables are separable, so we can easily find the closed forms for each variable. To optimize w.r.t π_0, we build the Lagrangian functional which induces only terms and constraints depending on π_0 :

$$\mathcal{L}(\pi_0, \lambda) = \sum_{i=1}^{K} p(q_0 = i|u;\theta^{k-1}) \log((\pi_0)_i) + \lambda(\sum_{i=1}^{K}(\pi_0)_i - 1)$$

We derive the expression, and we obtain (for all i) $\frac{p(q_0=i|u;\theta^{k-1})}{\pi_{0i}} + \lambda = 0$.
Using $\sum_{i=1}^{K} \pi_0 = 1$ and $\sum_{i=1}^{K} p(q_0 = i|u;\theta^{k-1}) = 1$, we result in computing the Lagrange multiplier $\lambda = -1$.

Therefore :

$$(\pi_0)_i = p(q_0 = i|u;\theta^{k-1})$$

To find a, we write the following Lagrangian (terms of a) :

$$\mathcal{L}(a, \mu) = \sum_{t=0}^{T-1}\sum_{i,j=1}^{K} p(q_{t+1} = i, q_t = j|y;\theta^{k-1}) \log(a_{i,j}) + \sum_{i=1}^{K} \mu_i(\sum_{j=1}^{K} a_{i,j} - 1)$$

We derive and obtain for all i,j :

$$\frac{\sum_{t=0}^{T-1} p(q_{t+1} = i, q_t = j|y;\theta^{k-1})}{a_{i,j}} + \mu_i = 0$$

.
Using the constraint that a is a transition matrix ($\forall i,\ \sum_{j=1}^{K} a_{i,j} = 1$), we have that $\mu_i = -\sum_{t=0}^{T-1} p(q_{t+1} = i|y;\theta^{k-1})$
Finally :

$$a_{i,j} = \frac{\sum_{t=0}^{T-1} p(q_{t+1} = i, q_t = j|y;\theta^{k-1})}{\sum_{t=0}^{T-1} p(q_{t+1} = i|y;\theta^{k-1})}$$

Now, we need to exhibit iterative estimations for μ and Σ. To optimize w.r.t the μ_i for all i, we just need to optimize the following functions :

$$g_i(\mu_i) = \frac{1}{2}(\sum_{t=0}^{T} p(q_t = i|y;\theta^{k-1}))(-\mu_i^T\Sigma_i^{-1}\mu_i) + \mu_i^T\Sigma^{-1}(\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})u_t$$

Now, writing that the gradient should be equal to 0 brings the following equality

$$\sum_{t=0}^{T} p(q_t = i|y;\theta^{k-1}))(\Sigma_i^{-1}\mu_i) = \Sigma_i^{-1}(\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})u_t$$

From this equality, we deduce the estimation of μ_i :

$$\mu_i = \frac{\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})u_t}{\sum_{t=0}^{T} p(q_t = i|y;\theta^{k-1})}$$

Concerning Σ, for all i, we minimize the following function (for the likelihood to be maximized w.r.t. Σ) :

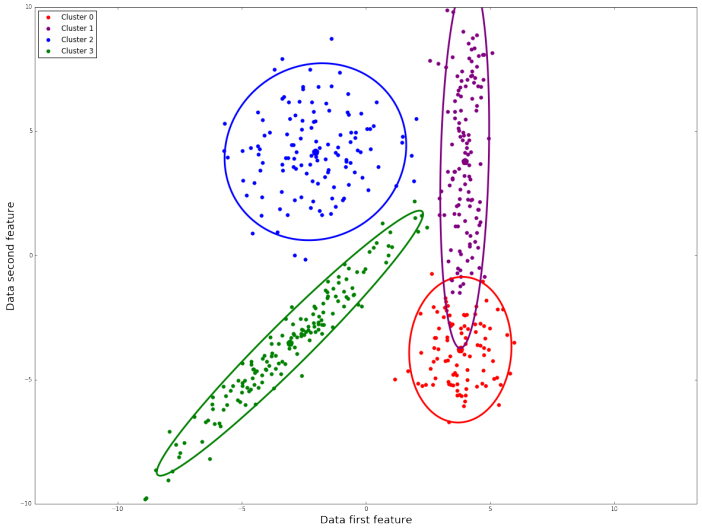
$$h_i(\Sigma_i) = \frac{\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})}{2} \log(|\Sigma_i|) + \frac{1}{2} Tr(\Sigma_i^{-1}(\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})(u_t - \mu_i)(u_t - \mu_i)^T))$$

.

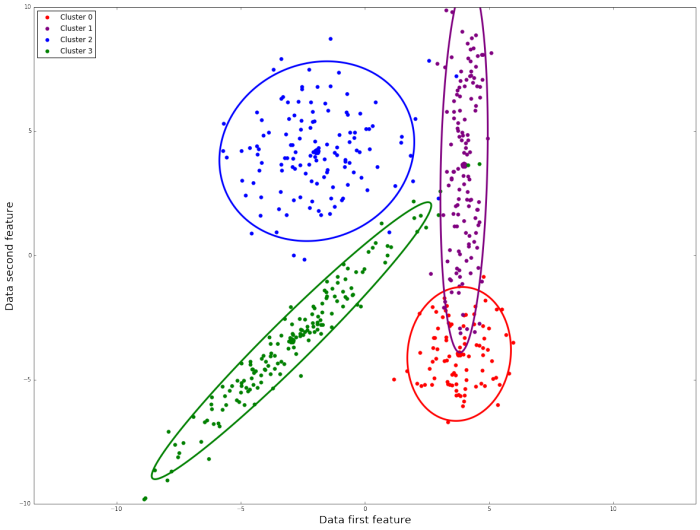
Using the classical property that $\nabla(\log(|X|)) = X^{-1}$ on $\mathcal{S}_n$, we then have that the gradient is equal to 0 when :

$$\Sigma_i = \frac{\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})(u_t - \mu_i)(u_t - \mu_i)^T}{\sum_{t=0}^{T-1} p(q_t = i|y;\theta^{k-1})}$$

Question 4



GMM clustering using EM algorithm (train, $\log \mathcal{L} = -4.74$)



HMM clustering using EM and inference algorithm (train, $\log \mathcal{L} = -3.81$)