Ecole Nationale Supérieur Paris-Saclay
Master Mathématiques Vision Apprentissage

Homework 3

DUCHEMIN Quentin & OREISTEIN Pierre

Cours: Probabilistic Graphical Models

December 2018

Implémentation - HMM

Question 1.2

Introduisons quelques notations liées au modèle HMM :

- $\forall k \in [\![1, K]\!], \quad \pi_k = p(q_0 = k)$
- $\forall (k, l) \in [\![1, K]\!]^2, \quad A_{k,l} = p(q_{t-1} = k \mid q_t = l)$ (pour t quelconque)
- $\forall t \in [\![0, T]\!], \ \forall k \in [\![1, K]\!], \quad \gamma_t^k = p(q_t = k \mid u_1, \dots, u_T)$
- $\forall t \in [\![1, T]\!], \ \forall (k, l) \in [\![1, K]\!]^2, \quad \gamma_{t-1,t}^{k,l} = p(q_{t-1} = k, q_t = l \mid u_1, \dots, u_T)$

L'indice i symbolise l'estimation obtenue à l'itération n°i de l'algorithme EM de la quantité correspondante. Une fois l'étape E effectuée et les probabilités a posteriori estimées, l'étape M à la $i-$ème itération s'effectue par :

$(\pi_k)^i$	$(A_{k,l})^i$	$(\mu_k)^i$	$(\Sigma_k)^i$
$(\gamma_0^k)^i$	$\frac{\sum_{t=1}^{T}(\gamma_{t-1,t}^{k,l})^i}{\sum_{t'=1}^{T}\sum_{l'=1}^{K}(\gamma_{t'-1,t'}^{k,l'})^i}$	$\frac{\sum_{t=0}^{T}(\gamma_t^k)^i u_t}{\sum_{t'=0}^{T}(\gamma_{t'}^k)^i}$	$\frac{\sum_{t=0}^{T}(\gamma_t^k)^i (u_t - \mu_k^i)(u_t - \mu_k^i)^T}{\sum_{t=0}^{T}(\gamma_t^k)^i}$

Pour les calculs qui ont permis l'obtention de ces résultats, on pourra se reporter à l'annexe.

Question 1.4

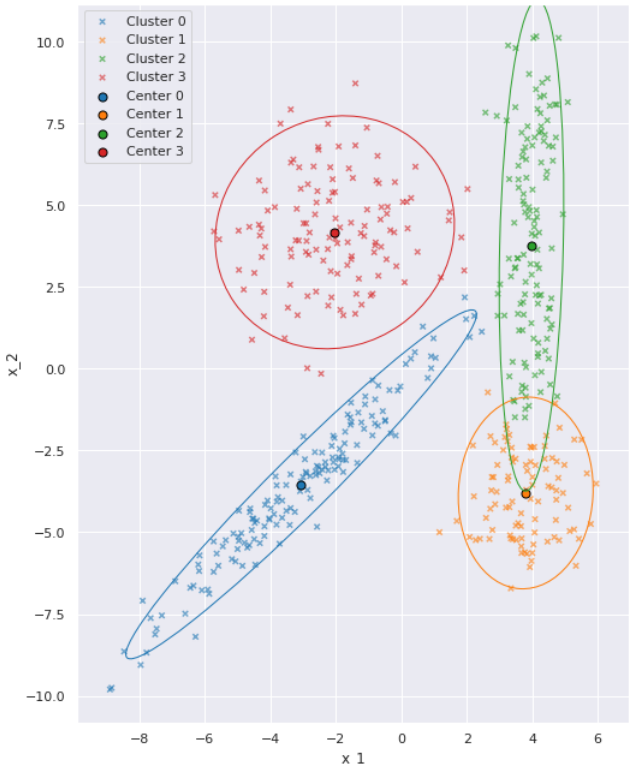

Figure 1 – Représentation des données d'entrainement. Les couleurs représentent le clustering donné par l'algorithm EM dans le cas du modèle GMM. Les ellipses contiennent 90% de la masse de la distribution de la gaussienne déterminée par l'EM.

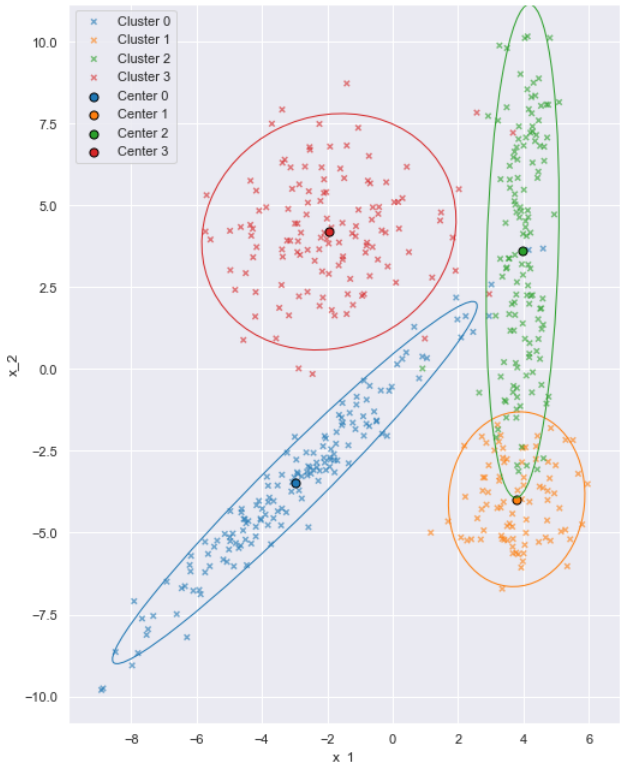

Figure 2 – Représentation des données d'entrainement. Les couleurs représentent le clustering donné par l'algorithme d'inférence pour le problème du decoding. Les ellipses contiennent 90% de la masse de la distribution de la gaussienne déterminée par l'EM précédement exécuté sur le modèle HMM.

Question 1.5

• Au vu des deux graphiques précédents on peut remarquer que les clusters sont très similaires. Ceci est assez logique vu que l'on a initialisé l'EM du modèle HMM avec les paramètres appris sur le modèle GMM. On peut toutefois noter quelques différences intéressantes. En particulier, les points qui appartiennent à l'intersection des ellipses verte et jaune différent dans leur classification. Dans le modèle GMM, ces points ont été classifiés comme "jaune" car le centroïde jaune était plus proche. A l'inverse, dans le modèle HMM, ces points ont été classifiés en "vert". Cela sous-entend donc qu'il existe bien une relation temporelle entre les variables u_t. Pour confirmer cette intuition, nous avons rajouter quelques lignes de code à notre notebook pour afficher récursivement les points sur le plan (cf section 5 du notebook). On peut alors constater qu'il existe bien une relation temporelle entre les u_t. En effet, on peut par exemple constater que l'apparition des points des clusters du haut s'alterne de manière régulière alors que les points des clusters du bas apparaissent en groupe les uns à la suite des autres.

• Suite à ces observations, il est naturel de se demander quel est le modèle qui permet d'atteindre la plus grande vraisemblance. Nous allons donc, pour les comparer, calculer la valeur de la log-vraisemblance (celle du modèle GMM) avec les paramètres appris pour les deux modèles. Ci-dessous nous présentons la valeur de cette log-vraisemblance normalisée par le nombre de données:

	Train	Test
GMM	$-1.30 * 10^{-5}$	$-9.63 * 10^{-10}$
HMM	$-6.51 * 10^{-13}$	$-2.00 * 10^{-18}$

Au vu de ces résultats, on constate donc que le modèle HMM performe mieux à la fois sur les données d'entraînement et de validation. Ceci confirme donc que ce n'est pas simplement du sur-apprentissage mais bien une composante des données. Les données répondent bien à une logique temporelle comme nous avons pu l'observer dans notre code (cf section 5 du notebook).

• Finalement, ces résultats sont assez naturels. Ils traduisent le fait que le modèle HMM permet de mieux appréhender la classification en donnant plus de poids aux clusters appropriés en prenant en compte cette composante temporelle. En effet, on peut constater que les formules de l'étape M sont presque inchangées entre les modèles HMM et GMM. Seules les probabilités à posteriori des classes sont modifiées. Ces résultats confirment donc que le modèle HMM attribue plus de poids (une plus grande probabilité) aux classes adaptées vu qu'il réussit à augmenter la vraisemblance. La prise en compte de la composante temporelle permet d'améliorer nos résultats de classification.

Annexes

Implémentation - HMM

Question 1.2

Introduisons quelques notations liées au modèle HMM :

- $\forall k \in [\![1, K]\!], \quad \pi_k = p(q_0 = k)$
- $\forall (k, l) \in [\![1, K]\!]^2, \quad A_{k,l} = p(q_{t-1} = k \mid q_t = l)$ (pour t quelconque)
- $\forall t \in [\![0, T]\!], \ \forall k \in [\![1, K]\!], \quad \gamma_t^k = p(q_t = k \mid u_1, \ldots, u_T)$
- $\forall t \in [\![1, T]\!], \ \forall (k, l) \in [\![1, K]\!]^2, \quad \gamma_{t-1,t}^{k,l} = p(q_{t-1} = k, q_t = l \mid u_1, \ldots, u_T)$

On se place dans un modèle HMM. La probabilité de l'événement $(u_1, ..., u_T, q_0, ..., q_T)$, se factorise sur le graphe de la manière suivante:

$$p(u_0, ..., u_T, q_0, ..., q_T) = p(q_0) \ * \prod_{t=1}^{T} p(q_t \mid q_{t-1}) \ * \prod_{t=0}^{T} p(u_t \mid q_t)$$

De cette formule, on déduit la complète log-vraisemblance :

$$\log(VS) = \log\left(p(q_0)\right) \ + \ \log\left(\prod_{t=1}^{T} p(q_t \mid q_{t-1})\right) \ + \ \log\left(\prod_{t=0}^{T} p(u_t \mid q_t)\right)$$

En introduisant les variables "one-hot encoding" $\delta_t^k = \begin{cases} 1 & \text{si } q_t = k \\ 0 & \text{sinon} \end{cases}$, et $\delta_t^{k,l} = \begin{cases} 1 & \text{si } q_{t-1} = k \text{ et } q_t = l \\ 0 & \text{sinon} \end{cases}$

on obtient:

$$\begin{aligned}
\log(VS) &= \log\left(\prod_{k=1}^{K} p(q_0 = k)^{\delta_0^k}\right) \ + \ \log\left(\prod_{t=1}^{T}\prod_{k,l=1}^{K} p(q_t = l \mid q_{t-1} = k)^{\delta_t^{k,l}}\right) \ + \ \log\left(\prod_{t=0}^{T}\prod_{k=1}^{K} p(u_t|q_t = k)^{\delta_t^k}\right) \\
&= \sum_{k=1}^{K} \delta_0^k \log(p(q_0 = k)) \ + \ \sum_{t=1}^{T}\sum_{k,l=1}^{K} \delta_t^{k,l} \log\left(\ p(q_t = l \mid q_{t-1} = k)\right) \ + \ \sum_{t=0}^{T}\sum_{k=1}^{K} \delta_t^k \log\left(p(u_t|q_t = k)\right)
\end{aligned}$$

• On se place à une itération i de l'algorithme EM. On suppose que l'étape E vient d'être réalisée, i.e. que nous venons de calculer les probabilités a posteriori:

$$\begin{aligned}
(\gamma_t^k)^i &= \mathbb{E}[\delta_t^k \mid u_{1,...,T}, \theta_{i-1}] \\
&= \mathbb{P}(q_t = k \mid u_{1,...,T}, \theta_{i-1}) \\
&= \frac{\alpha_t(k) * \beta_t(k)}{\sum_{l=1}^{K} \alpha_t(l) * \beta_t(l)}
\end{aligned}$$

Ainsi que les probabilités à posteriori :

$$\begin{aligned}
(\gamma_{t-1,t}^{k,l})^i &= \mathbb{E}[\delta_t^{k,l} \mid u_{1,...,T}, \theta_{i-1}] \\
&= \mathbb{P}(q_t = l, q_{t-1} = k \mid u_{1,...,T}, \theta_{i-1}) \\
&= \frac{\alpha_{t-1}(k) * \beta_t(l) * p(q_t = l \mid q_{t-1} = k) * p(u_t \mid q_t = l)}{\sum_{s=1}^{K} \alpha_t(s) * \beta_t(s)}
\end{aligned}$$

où nous avons noté $\theta_{i-1} = (\bar{a}, (\bar{\pi}_k, \bar{\mu}_k, \bar{\Sigma}_k)_{k\in[\![1,K]\!]})$; les paramètres appris via l'étape M de l'itération $i-1$ de l'algorithme.

• L'étape M consiste alors à maximiser l'espérance de la log-vraisemblance complète sachant les observations $(u_t)_{t=0...T}$ et les probabilités $(\gamma_t^k)^i$ et $(\gamma_{t-1,t}^{k,l})^i$.

$$\mathbb{E}[\ \log(VS) \mid (u_t)_{t=0...T}] = \sum_k^K (\gamma_0^k)^i \log(p(q_0 = k)) \ + \ \sum_{t=1}^T \sum_{k,l}^K (\gamma_{t-1,t}^{k,l})^i \log\left(p(q_t = l \mid q_{t-1} = k)\right) + \ \sum_{t=0}^T \sum_k^K (\gamma_t^k)^i \log\left(p(u_t \mid q_t = k)\right)$$

En utilisant les notations de l'énoncé, on obtient:

$$\mathbb{E}[\ \log(VS) \mid (u_t)_{t=0...T}] = \sum_k^K (\gamma_0^k)^i \log(\pi_k) \ + \ \sum_{t=1}^T \sum_{k,l}^K (\gamma_{t-1,t}^{k,l})^i \log\left(A_{k,l}\right) + \ \sum_{t=0}^T \sum_k^K (\gamma_t^k)^i \left(\frac{-1}{2} \log((2\pi)^d |\Sigma_k^i|) - \frac{(u_t - \mu_k^i)^T (\Sigma_k^i)^{-1} (u_t - \mu_k^i)}{2} \right)$$

La problème d'optimisation étant séparable en k et k, l, on va donc maximiser pour tout $k, l \in [\![1, K]\!]$. En particulier, nous cherchons donc à maximiser sur un ouvert convexe non vide une fonction concave. La condition nécessaire et suffisante d'optimalité s'écrit alors en annulant le gradient du lagrangien de notre problème. Nous obtenons alors :

- En dérivant par rapport à μ_k^i:

$$\sum_{t=1}^T (\gamma_t^k)^i \ \Sigma_k^{-1} \ (u_t - \mu_k^i) = 0$$

$$\boxed{\mu_k^i = \frac{\sum_{t=1}^T (\gamma_t^k)^i \ u_t}{\sum_{t'=1}^T (\gamma_{t'}^k)^i}}$$

- En dérivant par rapport à Σ_k^i:

$$\sum_{t=1}^T -\frac{(\gamma_t^k)^i}{2} \Sigma_k^i + \frac{(\gamma_t^k)^i}{2} (u_t - \mu_k^i)(u_t - \mu_k^i)^T = 0$$

$$\boxed{\Sigma_k^i = \frac{\sum_{t=1}^T (\gamma_t^k)^i (u_t - \mu_k^i)(u_t - \mu_k^i)^T}{\sum_{t'=1}^T (\gamma_{t'}^k)^i}}$$

- En dérivant par rapport à π_k le lagrangien:

  Notant λ le multiplicateur de Lagrange associé à la contrainte $\sum_{k=1}^{K} \pi_k = 1$, on a :

$$\frac{(\gamma_0^k)^i}{\pi_k} - \lambda = 0$$
$$\lambda \ \pi_k = (\gamma_t^k)^i$$

  Les contraintes $\sum_{k=1}^{K} \pi_k = 1$ et $\sum_{k=1}^{K} \gamma_0^k = 1$ fournissent alors la relation $\lambda = 1$, d'où:

$$\boxed{\pi_k = (\gamma_0^k)^i}$$

- En dérivant par rapport à $A_{k,l}$ le lagrangien:

  Notant λ_k le multiplicateur de Lagrange associé à la contrainte $\sum_{l=1}^{K} A_{k,l} = 1$, on a :

$$\sum_{t=1}^{T} \frac{(\gamma_{t-1,t}^{k,l})^i}{A_{k,l}} - \lambda_k = 0$$
$$\lambda_k \ A_{k,l} = \sum_{t=1}^{T} (\gamma_{t-1,t}^{k,l})^i$$

  La contrainte $\sum_{l=1}^{K} A_{k,l} = 1$ fournit alors la relation $\lambda_k = \sum_{t=1}^{T} \ \sum_{l=1}^{K} (\gamma_{t-1,t}^{k,l})^i$, d'où:

$$\boxed{A_{k,l} = \frac{\sum_{t=1}^{T} (\gamma_{t-1,t}^{k,l})^i}{\sum_{t'=1}^{T} \ \sum_{l=1}^{K} (\gamma_{t'-1,t'}^{k,l})^i}}$$