Université de Cergy-Pontoise,
UFR Sciences & Techniques,
Département des sciences informatiques,
Master 1 I&ISC

Intelligence Artificielle

TD 2 : Expectation Maximization

Louis Annabi

23 février 2021

Le but de ce TP est de mettre en oeuvre l'algorithme Expectation Maximization (EM) pour estimer les paramètres d'un modèle de mélange gaussien (Gaussian Mixture Model ou GMM en anglais).

1 Rappel du cours

1.1 Modèle de mélange gaussien

Notre modèle comporte des variables *cachées*. Les variables cachées, en opposition aux variables observables/-visibles, sont des variables aléatoires dont les valeurs ne sont pas données dans le data set.

Dans le cadre de ce TP, on va s'intéresser au cas d'un modèle graphique simple, où la variable cachée est une variable aléatoire discrète, et la variable visible suit une loi normale dont les paramètres dépendent de la valeur que prend la variable cachée.

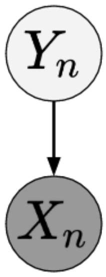

FIGURE 1 – Modèle graphique correspondant au modèle étudié dans ce TP. On représente en clair la variable aléatoire cachée, et en foncé la variable aléatoire observable.

La variable aléatoire discrète est paramétrée par K coefficients $(\pi_k)_{k=1...K}$ tels que pour toute valeur y_k possible, $p(Y = y_k) = \pi_k$. À chaque valeur possible de la variable aléatoire cachée Y correspond une loi de probabilité $p(X|Y = y_k)$. Dans le cas du mélange gaussien, on a :

$$\begin{aligned} p(X = x|Y = y_k) &= \mathcal{N}(x; \mu_k, \Sigma_k) && (1) \\ &= \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\Big\{ -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)\Big\} && (2) \end{aligned}$$

où la notation $|\Sigma|$ signifie le déterminant de la matrice de covariance Σ et où d correspond à la dimension de la variable aléatoire observable X.

On introduit également la notation τ_k pour désigner la probabilité a posteriori (c'est à dire en tenant compte de la variable observable) que $Y = y_k$, que l'on peut exprimer de la manière suivante en utilisant la loi de Bayes :

$$\begin{aligned}
\tau_k &= p(Y = y_k|x) && (3)\\
&= \frac{p(X = x|Y = y_k)p(Y = y_k)}{p(X = x)} && (4)\\
&= \frac{\pi_k \mathcal{N}(x; \mu_k, \Sigma_k)}{\sum_{k'} \pi_{k'} \mathcal{N}(x; \mu_{k'}, \Sigma_{k'})} && (5)
\end{aligned}$$

1.2 Estimation des paramètres

On s'intéresse au problème d'estimer les paramètres $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1\ldots K}$ de notre modèle à partir des données observables $(x_i)_{i=1\ldots n}$. Dans ce TP, ces données observables seront un ensemble de points en 2D.

Pour estimer les paramètres, on souhaiterait employer la méthode (courante en statistiques) du maximum de vraisemblance. Cette méthode calcule la vraisemblance d'un jeu de données en fonction des paramètres θ et choisit la valeur de θ qui maximise cette quantité. Dans notre cas, le calcul de la vraisemblance donne :

$$\begin{aligned}
\mathcal{L}(\theta) &= \sum_{i=1}^{n} \log p(x_i; \theta) && (6)\\
&= \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \mathcal{N}(x_i; \mu_k, \Sigma_k) && (7)
\end{aligned}$$

Il n'est pas possible de trouver la solution algébrique qui maximise cette quantité, à la place nous allons donc employer l'algorithme EM, qui cherche à maximiser cette quantité en modifiant itérativement une estimation des paramètres $\theta^{(t)}$ et une estimation des probabilité a posteriori $\tau_k^{(t)}$.

1.3 Algorithme EM

Cet algorithme alterne itérativement deux phases :

Expectation : Dans cette phase, on estime à partir de notre estimation des paramètres $\theta^{(t)}$, la probabilité a posteriori de la variable cachée Y pour chaque donnée :

$$\tau_{k,i}^{(t)} = \frac{\pi_k^{(t)} \mathcal{N}(x_i; \mu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{k'=1}^{K} \pi_{k'}^{(t)} \mathcal{N}(x_i; \mu_{k'}^{(t)}, \Sigma_{k'}^{(t)})} \tag{8}$$

Maximization : Dans cette phase, on met à jour notre estimation des paramètres en utilisant l'estimation de la probabilité a posteriori $\tau_{k,i}^{(t)}$. Nous allons voir comment réaliser cette estimation dans un cas simplifié et dans le cas général.

L'idée derrière cet algorithme (qu'on ne démontrera pas ici) est qu'on a une garantie que ce processus itératif fait augmenter la vraisemblance $\mathcal{L}(\theta)$. En revanche, il n'y a pas de garantie que la solution vers laquelle cet algorithme converge est la solution optimale.

1.3.1 Cas simplifié

Dans les deux cas, l'estimation des paramètres $\pi_k^{(t+1)}$ et $\mu_k^{(t+1)}$ est la même :

$$\pi_k^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_{k,i}^{(t)} \tag{9}$$

$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{k,i}^{(t)} x_i}{\sum_{i=1}^{n} \tau_{k,i}^{(t)}} \tag{10}$$

Dans le cas simplifié, on contraint les matrices de covariance à être de la forme $\Sigma_k = \mathbb{I}_2$ où $\mathbb{I}_2$ est la matrice identité de dimension 2. En conséquence, il n'y a pas de paramètre supplémentaire à estimer.

1.3.2 Cas général

Dans le cas général, on estime également les paramètres des matrices de covariance :

$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} \tau_{k,i}^{(t)} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_{i=1}^{n} \tau_{k,i}^{(t)}} \tag{11}$$

2 Travail demandé

Le data set qui vous est fourni comporte 500 points de dimension 2. Pour l'implémentation de l'algorithme, on utilisera $K = 3$.

- Implémentez une fonction calculant la probabilité $\mathcal{N}(x; \mu, \Sigma)$ (équation 2)
- Implémentez une fonction calculant la vraisemblance $\mathcal{L}(\theta)$ pour un jeu de données en fonction des paramètres $\theta = (\pi_k, \mu_k, \Sigma_k)_{k=1\ldots K}$ (équation 7)
- Implémentez une fonction affichant dans un graphique votre nuage de points, et les ellipses correspondant aux trois gaussiennes
- Implémentez une fonction correspondant à la phase Expectation (équation 8)
- Implémentez une fonction correspondant à la phase Maximization (équations 9,10,11)
- Implémentez l'algorithme EM en alternant les phases E et M
- Générez un graphique suivant l'évolution de la vraisemblance au cours de l'apprentissage

Une fois l'implémentation terminée, essayez les deux configurations : cas simplifié et cas général, et comparez vos résultats (affichage des gaussiennes et vraisemblance).

- Quelle configuration fonctionne le mieux en terme de vraisemblance et au niveau visuel ?
- D'après vous, si on augmentait le nombre de gaussiennes, par exemple en prenant $K = 4$, la vraisemblance obtenue devrait elle augmenter ou diminuer ? Justifiez votre réponse.

3 Rendu

Votre code et un rapport de 2 à 5 pages contenant les réponses aux questions, les affichages demandés ainsi que vos analyses sont à rendre le week-end suivant le TP.