Application de l'algorithme EM à un mélange de Gaussiennes

Tom CIFFREO, Petr ZAMOLODTCHIKOV

ABSTRACT

L'objectif de ce travail est d'implémenter l'algorithme EM dans le cadre d'un mélange de gaussiennes dans un modèle en information incomplète

Préliminaires

Question 1:

Soit Z une variable aléatoire à valeurs dans $1,...,K$ telle que $P(Z_i = k) = P(X_i \sim f_{\theta_k})$.
Nous travaillons dans le cadre d'un modèle à données manquantes (symbolisé par la variable latente Z).
Ainsi, la vraisemblance du modèle s'écrit:

$$L(\mathrm{X}, \theta) = \int_{\Omega} L(\mathrm{X}, z; \theta)\, \mathrm{d}\mu(z)$$

C'est à dire que la densité de mélange (ie la densité de X) s'écrit:

$$\begin{aligned} f(x) &= \int_{\{1,...,K\}} f(x, z; \theta)\, \mathrm{d}\mu(z) && (1)\\ &= \int_{\{1,...,K\}} f(x|z; \theta) f(z)\, \mathrm{d}\mu(z) && (2)\\ &= \int_{\{1,...,K\}} f_{\theta_k}(x) f(z)\, \mathrm{d}\mu(z) && (3)\\ &= \sum_{k=1}^{K} f_{\theta_k}(x) P(Z_i = k) && (4)\\ &= \sum_{k=1}^{K} \alpha_k f_{\theta_k}(x) && (5) \end{aligned}$$

Les α_k représentent les poids de chaque densité ie le nombre de variables suivant chaque loi caractérisée par la densité f_{θ_k}.
On retrouve bien la propriété d'une densité de mélange, comme combinaison convexe des densités des variables du modèle où la somme des poids de chaque densité vaut 1.

Question 2:

Dans un problème supervisé, on a à notre disposition les données nous indiquant l'appartenance de chaque variable X_i à une classe. Dans un modèle non supervisé, cette étiquette manque: il nous manque une partie des données. De manière générale, dans un problème supervisé, on connaît les k classes, contrairement à un problème non supervisé. Ainsi, nous sommes ici dans le cadre d'un problème d'apprentissage non supervisé puisque l'on ne connaît pas l'appartenance des variables aux classes (symbolisé par les variables latentes Z_i). Il s'agit donc d'un problème de classification non supervisé ou clustering (on veut répartir les X_i selon les données que l'on observe en différentes classes, qui correspondent aux différentes densités).

1 Mélange de gaussiennes et algorithme EM

Question 3:

On a $\theta_k = (\mu_k, \Sigma_k)$ et par définition, la vraisemblance s'écrit:

$$\begin{aligned} L(X;\theta) &= \prod_{i=1}^{n} f_{X_i}(x_i;\theta) \\ &= \prod_{i=1}^{n}\sum_{k=1}^{K} \alpha_k f_{\theta_k}(x_i) \\ L(X;\theta) &= \prod_{i=1}^{n}\sum_{k=1}^{K} \frac{\alpha_k}{(2\pi)^{p/2} det(\sigma_k^2)^{1/2}} \exp(-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k)) \end{aligned}$$

Si $p = 1$, on a $\theta_k = (\mu_k, \sigma_k^2) \in \mathbb{R} \times \mathbb{R}$:

$$\begin{aligned} L(X;\theta) &= \prod_{i=1}^{n}\sum_{k=1}^{K} \frac{\alpha_k}{(2\pi)^{1/2}(\sigma_k^2)^{1/2}} \exp(-\frac{(x_i-\mu_k)^2}{2\sigma_k^2}) \\ L(X;\theta) &= \sum_{k=1}^{K} (\frac{\alpha_k}{(2\pi)^{1/2}(\sigma_k^2)^{1/2}})^n \exp(-\sum i = 1^n \frac{(x_i-\mu_k)^2}{2\sigma_k^2}) \end{aligned}$$

Si $p = 2$, on a $\theta_k = (\mu_k, \Sigma_k) \in \mathbb{R}^2 \times \mathbb{R}^{2\times 2}$:

$$\begin{aligned} L(X;\theta) &= \prod_{i=1}^{n}\sum_{k=1}^{K} \frac{\alpha_k}{2\pi(\det(\Sigma_k))^{1/2}} \exp(-\frac{1}{2}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)) \\ L(X;\theta) &= \sum_{k=1}^{K} (\frac{\alpha_k}{2\pi(\det(\Sigma_k))^{1/2}})^n \exp(-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_k)^T \Sigma_k^{-1}(x_i-\mu_k)) \end{aligned}$$

Question 4:

L'estimateur du Maximum de Vraisemblance n'est pas adapté ici car dès que p est plus grand que 1, il s'agit d'optimiser la vraisemblance sur un grand nombre de paramètres, ce qui est difficile à faire de manière analytique. De plus, l'EMV est principalement utilisé dans le cadre d'un problème supervisé, où l'on connait les paramètres, ce qui n'est pas le cas dans notre modèle. L'intérêt de l'algorithme EM ici est qu'il procède par itérations (ce qui permet de palier au problème de données manquantes en réestimant ces dernières à chaque itération) et qu'il est plus adapté pour maximiser une expression complexe sur un grand nombre de paramètres.

Question 5:

L'algorithme EM (Expectation-Maximization) permet de calculer l'estimateur du maximum de vraisemblance (EMV) par itérations dans le cadre d'un modèle à données manquantes. Il consiste en deux étapes (E-step et M-step) que l'on réitère afin de converger vers l'EMV. L'algorithme procède comme suit:

Etape initiale: on choisit de manière arbitraire un $\theta^{(0)}$ pour initialiser l'algorithme et pour tout $t \geq 1$, on répète les étapes 1 et 2.

Etape 1: (Expectation Step) on calcule l'Espérance de la vraisemblance complète du modèle $L(X,Z;\theta)$ sous la loi de Z sachant $X = x$ et $\theta^{(t)}$:

$$\mathbb{E}_{Z|X=x,\theta^{(t)}}[L(X,Z,\theta)]$$

Etape 2: (Maximization Step) on maximise l'espérance sur $\theta \in \Theta$ pour obtenir $\theta^{(t+1)}$:

$$\theta^{(t+1)} \in \underset{\theta\in\Theta}{\operatorname{argmax}}\{\mathbb{E}_{Z|X=x,\theta^{(t)}}[L(X,Z,\theta)]\}$$

Il faut également se souvenir de l'expression de la vraisemblance complète:

$$L(X,Z,\theta) = \prod_{i=1}^{n} f(x_i,z_i,\theta) = \prod_{i=1}^{n} f(z_i|x_i,\theta) f(x_i|\theta)$$

L'algorithme EM relie à chaque étape le problème en données manquantes à un problème à données complétées sur lequel on estime le θ. On peut alors réestimer les données manquantes et recalculer le θ de sorte qu'à chaque incrémentation l'estimation est de plus en plus précise car l'estimation des données manquantes est de plus en plus pertinente. Ceci se traduit par le fait que la vraisemblance augmente à chaque incrémentation de l'algorithme:

$$\forall t \geq 0 : L(X,Z;\theta^{(t+1)}) \geq L(X,Z;\theta^{(t)})$$

Question 6:

L'algorithme du K-means (ou K-moyennes) s'utilise dans des problèmes de classification non supervisée. Il permet de séparer en k clusters (k déterminé en initialisant l'algorithme) en minimisant la distance d'un point à la moyenne des points de son cluster:

$$\min \sum_{i,j} d_{ij}^2$$

On l'utilise avant un algorithme EM afin de choisir une valeur $\theta^{(0)}$ pertinente car la convergence de l'algorithme EM vers l'EMV dépend de la valeur d'initialisation. En fait, le K-means permet de remplir à l'étape initiale les données manquantes du modèle afin de se ramener à un modèle en données complètes et ainsi de choisir simplement le $\theta^{(0)}$. Il procède de la manière suivante:

1. On choisit k individus représentant les barycentres des k classes que l'on cherche.
2. On calcule la distance de chaque point aux différents barycentres et on affecte à chaque point la classe dont le barycentre est le plus proche.
3. On recalcule les barycentres des nouvelles classes obtenues à l'étape 2.
4. On réitère les étapes précédentes jusqu'à la stabilisation des barycentres.

Question 7:

L'algorithme EM converge (car la vraisemblance augmente à chaque étape. Or celle-ci est convexe bornée comme produit de fonctions de densité bornées) mais pas nécessairement vers un optimum global (ie il ne converge pas nécessairement vers l'EMV). La convergence vers un optimum global dépend de la valeur $\theta^{(0)}$ à laquelle on initialise l'algorithme. Il peut en effet converger vers un maximum local. C'est précisément pour s'assurer de la convergence vers le maximum global que l'on execute dans un premier temps un K-means qui lui converge assurément (car il consiste juste en un clustering de données).

2 Implémentation de l'algorithme EM

Question 8:

L'implémentation de l'algorithme K-means se trouve dans le code joint.

Après application du K-means au jeu de données: L'algorithme K-means est sensible aux centroïdes initiaux, en particulier pour de grandes dimensions. Ici, en dimension 2, l'algorithme converge sensiblement vers les mêmes clusters à chaque fois.

Question 9:

On définit les parts de chaque point dans chaque distibution:

$$\gamma_{i,k}^{t+1} = \frac{\alpha_k^t f_k(x_i|\mu_k^t,\Sigma_k^t)}{\sum_{j=1}^{N} \alpha_j^t f_j(x_i|\mu_j^t,\Sigma_j^t)}$$

Soit $S_k^{t+1} = \sum_{i=1}^{N} \gamma_{i,k}^{t+1}$

L'équation associée à la M-step s'écrit dans le cas général :

$$\alpha_k^{t+1} = \frac{1}{N} S_k^{t+1}$$

$$\mu_k^{t+1} = \frac{1}{S_k^{t+1}} \sum_{i=1}^{N} \gamma_{i,k}^{t+1}$$

$$\sigma_{l,m,k}^{t+1} = \sqrt{\frac{1}{S_k^{t+1}} \sum_{i=1}^{N} \gamma_{i,k}^{t+1} (X_{l,i} - \mu_{l,k}^{t+1})(X_{m,i} - \mu_{m,k}^{t+1})}$$

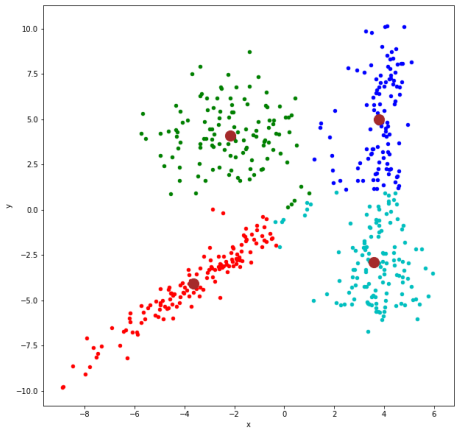

Figure 1. K-means clustering. En marron: les centres de gravité obtenus

Avec la supposition de diagonalité des matrices de covariance, nous avons appliqué:
$\Sigma_k^{t+1} = (\sigma_{1,1,k}^{t+1})^2 I_2$
Les clusters sont obtenus en assignant chaque point x_i au couple (μ_k, Σ_k) maximisant $f_{\theta_k(x_i)}$

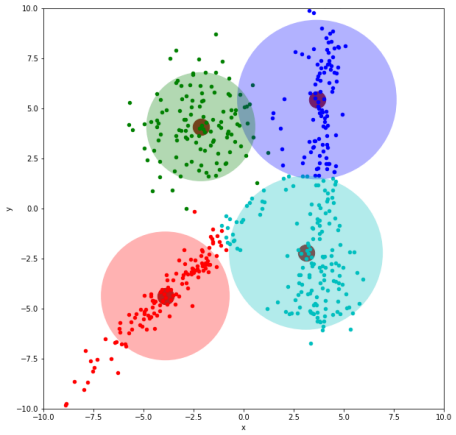

Figure 2. EM matrices de variance diagonales.En marron; les centres de gravité obtenus, les ellipses correspondent à la ligne de niveau $f_{\theta_k}(x) = 0.8$

Question 10:
Le cas général est traité dans la question précédente; la figure 3 représente le clustering obtenu.

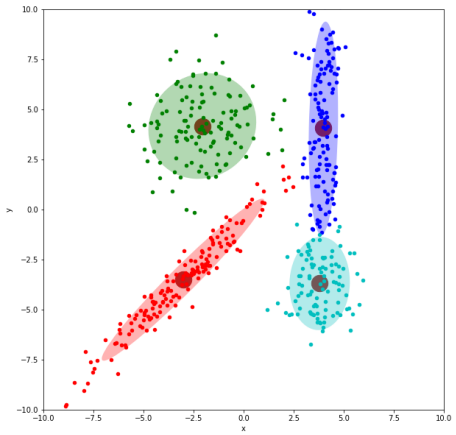

Figure 3. EM sans hypothèses sur la forme des matrices de variance.En marron; les centres de gravité obtenus, les ellipses correspondent à la ligne de niveau $f_{\theta_k}(x) = 0.8$

Question bonus:
Les ellipses ont étées tracées sur tous les graphiques, on voit qu'en supposant $\Sigma_k = \sigma_k^2 I_2$ les ellipses sont des cercles (ce qui correspond bien au cas ou les composantes du vecteur gaussien son décorélées). Dans le cas général les ellipses épousent bien mieux les données et les clustering est plus précis.

Question 11:
Le modèle $\Sigma_k = \sigma_k^2 I_2$ n'est pas adapté car :

- Dans notre problème les données ont deux coordonnées, on peut donc les visualiser aisément et l'on voit bien que les clusters ne ressemblent pas à des cercles, en faisant cette supposition on trouve donc des matrices de variance fausses et il n'y a aucune chance de converger vers les paramètres reels.
- Même si il y avait plus de coordonnées, l'hypothèse devrait être vérifiée en calculant les vocariance empiriques des données et tester l'hypothèse nulle avec un test du chi-deux par exemple. Ici l'hypothèse nulle serait rejetée.

Finalement supposer les matrices de covariance nulles peut, moyennant vérification rendre l'algorithme plus rapide, mais, dans notre cas, ce modèle n'est pas adapté.