TD 6

Exercice 1 – Algorithme des K-moyennes

La taille du dictionnaire K est fixée, c'est un paramètre de l'algorithme.

- Initialiser aléatoiremet les K prototypes.
- Répéter jusqu'à (critère d'arrêt) :
  - ▶ partitionner les exemples en les affectant aux prtotypes dont ils sont le plus proche ;
  - ▶ redéfinir les protoypes (i.e. centres de gravité des partitions).

Q 1.1 Soit l'ensemble d'exemples en dimension 2 :
$D = \{(0,-4),(0,-3),(1,-3),(1,-2),(0,4),(-1,1),(-1,2),(0,3)\}$
Faire tourner l'algorithme des K-moyennes en prenant comme point de départ les prototypes $(0,-6)$ et $(-1,1)$.

1iere iteration clusters : $\{1,2,3\}$ et $\{4,5,6,7,8\}$. Prototypes : $(0.33,-3.33)$ et $(-1/5,8/5)$
2ième iteration clusters : $\{1,2,3,4\}$ et $\{5,6,7,8\}$. Prototypes : $(0.5,-3)$ et $(-0.5,2.5)$
3ième iteration = clusters 2ième iteration. Fin

Q 1.2 Quels critères d'arrêt préconisez-vous pour les méthodes de QV ?

K-Means = stabilité des clusters

Exercice 2 – Clustering et mélange de lois

On souhaite estimer une densité de probabilité par un modèle de type mélange de gaussiennes. La probabilité d'une observation x est donnée par : $p(x) = \sum_{l=1}^{L} \tau_l.p(x|\lambda_l)$ où les τ_l sont les probabilités a priori des lois et les $p(x|\lambda_l)$ sont des lois gaussiennes multi-dimentionnelles caractérisées par leur moyenne μ_l et leur matrice de co-variance Σ_l , i.e. $\lambda_l = (\mu_l, \Sigma_l)$.

Q 2.1 Dessiner la loi de probabilité pour $L = 2, \tau_1 = \tau_2 = 0.5$, et $\mu_1 = 1, \mu_2 = 3, \Sigma_1 = 1, \Sigma_2 = 10$.

Q 2.2 Quelles est la probabilité a posteriori qu'un exemple x ait été produit par la gaussienne multi-dimentionnelle l, $p(\lambda_l|x)$?

$$p(\lambda_l|x) = \frac{p(x|\lambda_l)\tau_l}{\sum_{l=1}^{L} \tau_l.p(x|\lambda_l)}$$
Avec $p(x|\lambda_l) = \frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}_l|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^{\mathrm{T}}\mathbf{\Sigma}_l{}^{-1}(\mathbf{x}-\boldsymbol{\mu}_l)\right)$,

Q 2.3 Expliquer comment l'apprentissage d'un mélange de lois peut être utilisé pour faire du clustering.

Si l'on a tous les paramètres d'un mélange, on peut estimer les probabilités a posteriori d'appartenance des x à chaque cluster selon $p(\lambda_l|x)$ la proba qu'il ait été produit par la gaussienne correspondante

Exercice 3 – Apprentisage d'un mélange de lois et maximum de vraissemblance

On souhaite apprendre le modèle de l'exercice précedent avec un critère de maximum de vraisemblance (MV) sur une base d'apprentissage $X = \{x_i\}, i = 1..N$.

Q 3.1 Exprimer la log-vraisemblance des données par le modèle en supposant que les x_i sont indépendants.

Soit θ l'ensemble des paramètres des différentes gaussiennes λ_l et leurs proportions P_l
$L(\theta, X) = \prod_{x_i \in X} p(x_i|\theta) = \prod_{x_i \in X} \sum_{l=1}^{L} \tau_l p(x_i|\lambda_l)$
Puisque log strictement croissante revient à maximiser :
$\log L(\theta, X) = \sum_{x_i \in X} \log p(x_i|\theta) = \sum_{x_i \in X} \log \sum_{l=1}^{L} \tau_l p(x_i|\lambda_l)$

Q 3.2 Quelle est la difficulté avec cette log-vraissemblance ?

On a une somme de log d'une somme, difficile à optimiser

Q 3.3 L'idée de l'algorithme EM (Expectation-Maximization) est de se dire que si l'on avait des informations supplémentaires Z, on pourrait optimiser cette vraissemblance plus facilement. Quelles informations seraient utiles ici ? Donner la vraissemblance complétée par ces informations.

Z=Les affectations des éléments aux différents clusters
$\log L(\theta; x, z) = \log p(X, Z|\theta) = \sum_{x_i} \log \sum_{j=1}^{L} \mathbb{I}(z_i = j)\ p(x_i|\lambda_j)\tau_j$
ou
$\log L(\theta; x, z) = \sum_{x_i} \sum_{j=1}^{L} \mathbb{I}(z_i = j)\big[\log \tau_j - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x_i - \mu_j)^\top \Sigma_j^{-1}(x_i - \mu_j) - \frac{d}{2}\log(2\pi)\big]$ car $\mathbb{I}(z_i = j)$ est différent de 0 pour un seul j

Q 3.4 On peut alors utiliser un algorithme dit algorithme EM (Expectation-Maximization) pour l'estimation de ce mélange de gaussiennes. Une variante de l'algorithme EM est la suivante :

- initialiser les paramètres $(\tau_i, \mu_i, \sigma_i)_{i=1..L}$;
- Répéter :
  - ▶ déterminer pour chaque x_i la gaussienne qui l'a produit avec la plus grande vraissemblance : pour $i = 1..N$, $I(x_i) = argmax_l\ p(\lambda_l|x_i)$;
  - ▶ ré-estimer les paramètres des lois à partir des exemples qui lui ont été affectés : pour $l = 1..L$, ré-estimer λ_l à partir des $\{x_i \in E | I(x_i) = l\}$

Dans le cas où tous les τ_i sont égaux (equi-probabilité des gaussiennes) et où les matrices de covariance des lois sont fixées à l'identité, montrer que l'algorithme précédent est équivalent à un algorithme des K-Moyennes.

$\log p(\lambda_l|x_i) \propto -\frac{1}{2}(x_i - \mu_l)^\top (x_i - \mu_l)$. Choisir $I(x_i) = argmax_l\ p(\lambda_l|x_i)$ revient alors à choisir le cluster de centre le plus proche comme dans l'algo k-means

Q 3.5 L'algorithme précédent procède par affectations successives des éléments aux différents clusters. Quelle est la limite de ce genre d'approche ?

Risque de s'enfermer rapidement dans des optima locaux (solutions sous-optimales)
Affectations "dures" des éléments aux clusters alors que leur appartenance peut être très incertaine

Q 3.6 La version classique de l'algorithme EM travaille en deux étapes :

- Expectation step (E step) : Calcul de l'espérance de la log-vraissemblance en fonction des probabilités conditionnelles des données manquantes Z étant donné les observations X selon les estimations courantes des paramètres $\theta^{(t)}$:

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{Z|X,\theta^{(t)}}\left[\log L(\theta; X; Z)\right]$$

- Maximization step (M step) : Recherche des paramètres θ qui maximisent cette quantité :

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

Q 3.7 Sachant que d'après l'inégalité de Gibbs, $\sum_{i=1}^{n} p_i \log p_i \geq \sum_{i=1}^{n} p_i \log q_i$ pour toutes paires de distributions de probabilités p et q, montrer que la suite des vraissemblances $p(X|\theta^{(t)})$, selon les paramètres $\theta^{(t)}$ calculés à chaque étape de l'algorithme EM, est croissante.

On a pour tout Z : $\log p(\mathbf{X}|\boldsymbol{\theta}) = \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta})$
On peut alors écrire :

$$\log p(\mathbf{X}|\boldsymbol{\theta}) = \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}) - \sum_{\mathbf{Z}} p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^{(t)}) \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}) \quad (1)$$

$$= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}), \quad (2)$$

Cette equation est valide pour tout $\boldsymbol{\theta}$ incluant $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$. On peut donc noter : $\log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$
On peut alors s'intéresser au gain résultant du fait de considérer un nouveau $\boldsymbol{\theta}$ plutôt que les paramètres courants $\boldsymbol{\theta}^{(t)}$: $\log p(\mathbf{X}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) = Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)}) + H(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - H(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$
Selon l'ingélaité de Gibbs, on peut en déduire : $\log p(\mathbf{X}|\boldsymbol{\theta}) - \log p(\mathbf{X}|\boldsymbol{\theta}^{(t)}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$
Ce qui indique que choisir $\boldsymbol{\theta}$ de manière à améliorer $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ par rapport à $Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$ permet d'améliorer $\log p(\mathbf{X}|\boldsymbol{\theta})$ d'au moins autant.
La suite des vraissemblances définie suivant les paramètres calculés à chaque étape de l'algorithme EM est donc croissante.

Q 3.8 Donner la formulation de l'espérance $Q(\theta|\theta^{(t)})$ selon les paramètres courants $\theta^{(t)}$.

$$T_{j,i}^{(t)} := \mathrm{P}(Z_i = j|X_i = \mathbf{x}_i; \theta^{(t)}) = \frac{\tau_j^{(t)}\ p(\mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \Sigma_j^{(t)})}{\sum_{l=1}^{L} \tau_l^{(t)}\ p(\mathbf{x}_i; \boldsymbol{\mu}_l^{(t)}, \Sigma_l^{(t)})}$$

$$Q(\theta|\theta^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{L} P(Z_i = j|X_i = \mathbf{x}_i; \theta^{(t)}) \log L(\theta_j; \mathbf{x}_i, \mathbf{z}_i) \quad (3)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{L} T_{j,i}^{(t)} \left[\log \tau_j - \tfrac{1}{2}\log|\Sigma_j| - \tfrac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_j) - \tfrac{d}{2}\log(2\pi)\right] \quad (4)$$

Q 3.9 Donner alors les formulations des estimations des paramètres à l'itération t+1 selon les estimations à l'itération t

$$\boldsymbol{\tau}^{(t+1)} = \arg\max_{\boldsymbol{\tau}} Q(\theta|\theta^{(t)}) \tag{5}$$

$$= \arg\max_{\boldsymbol{\tau}} \left\{ \sum_{j=1}^{L} \left[\sum_{i=1}^{n} T_{j,i}^{(t)} \right] \log \tau_j \right\} \tag{6}$$

Selon le MLE pour la binomial distribution :

$$\tau_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)}}{\sum_{i=1}^{n} \sum_{l=1}^{L} T_{l,i}^{(t)}} = \frac{1}{n} \sum_{i=1}^{n} T_{j,i}^{(t)}$$

$$(\boldsymbol{\mu}_j^{(t+1)}, \Sigma_j^{(t+1)}) = \arg\max_{\boldsymbol{\mu}_j, \Sigma_j} Q(\theta|\theta^{(t)}) \tag{7}$$

$$= \arg\max_{\boldsymbol{\mu}_j, \Sigma_j} \sum_{i=1}^{n} T_{j,i}^{(t)} \left\{ -\tfrac{1}{2} \log |\Sigma_j| - \tfrac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_j)^\top \Sigma_j^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_j) \right\} \tag{8}$$

Selon le MLE pour la normal distribution :

$$\boldsymbol{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)} \mathbf{x}_i}{\sum_{i=1}^{n} T_{j,i}^{(t)}}$$

Et

$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)} (\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})(\mathbf{x}_i - \boldsymbol{\mu}_j^{(t+1)})^\top}{\sum_{i=1}^{n} T_{j,i}^{(t)}}$$