Dorian Gaillard

Homework 2 :

Exercise 1)

1) First we subdivide $[0;1]$ into n intervals of length p_i :

$I_1 = [0; p_1]$

$I_i = [\sum_{k=1}^{i-1} p_k ; \sum_{k=1}^{i} p_k]$ for $i \in [2; n-1]$

$I_n = [\sum_{k=0}^{n-1} p_k ; 1]$

then we sample $U \sim \mathcal{U}([0;1])$ and assign $X = x_i$ when $U \in I_i$.

Then $P(X = x_i) = P(U \in I_i)$
$= \text{length}(I_i)$
$= p_i$

Exercise 2).

1) Here $\theta = (\mu_1, \dots, \mu_m, \Sigma_1, \dots, \Sigma_m, \alpha_1, \dots, \alpha_m)$

The likelihood is then

$$\mathcal{L}(x_1, \dots, x_n; \theta) = \prod_{i=1}^{n} p_\theta(X_i = x_i)$$
$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} p(X_i = x_i | Z_i = j) P(Z_i = j) \right)$$
$$= \prod_{i=1}^{n} \left(\sum_{j=1}^{m} p(X_i = x_i | X \sim \mathcal{N}(\mu_j, \Sigma_j)) \alpha_j \right)$$

3) The EM algorithm is made of two phases :

* first, the complete likelihood (knowing the

latent variables z) is

$$\log \mathcal{L}(X, Z; \Theta) = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\log(\alpha_j) + \log p(X_i = x_i \mid Z_i = j)\right) \times \mathbb{1}_{Z_i = j}$$

Then $B(\Theta, \Theta_k) = \mathbb{E}\left[\log \mathcal{L}(X, Z; \Theta) \mid X, \Theta_k\right]$

$$= \sum_{i=1}^{n} \sum_{j=1}^{n} P(Z_i = j \mid X_i, \Theta_k) \left[\log(\alpha_j) + \log(p(X_i = x_i \mid Z_i = j))\right]$$

* Then we compute $\Theta_{k+1} \in \arg\max B(\Theta, \Theta_k)$.

* α_j^{k+1} is solution to the minimisation problem

$$\min_{\alpha_j} -B(\Theta, \Theta_k)$$

st $\alpha_j \geq 0$,
$\sum_{j=1}^{n} \alpha_j = 1$

Using Lagrange multipliers the KKT conditions give us: $\nabla\left(-B(\Theta, \Theta_k) + \lambda\left(\sum_{j=1}^{n} \alpha_j - 1\right) - \nu \alpha_j\right) = 0$

ie $$-\sum_{i=1}^{n} \frac{P(Z_i = j \mid X_i = x_i, \Theta_k)}{\alpha_j} + \lambda - \nu = 0$$

and $\nu \alpha_j = 0$

so $\nu = 0$

and $$\lambda \alpha_j = \sum_{i=1}^{n} P(Z_i = j \mid X_i = x_i, \Theta_k)$$

Summing on j gives $\lambda = n$

so $$\alpha_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} P(Z_i = j \mid X_i = x_i, \Theta_k)$$

Then knowing $p(X_i = x_i | z_i = j) = \dfrac{e^{-\frac{1}{2}(x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j)}}{\sqrt{(2\pi)^d |\Sigma_j|}}$

$$* \nabla_{\mu_j} B(\theta, \theta_k) = \sum_{i=1}^{n} P(z_i = j | X_i, \theta_k) (\Sigma_j^{-1})^T (x_i - \mu_j) = 0$$

$$\Longrightarrow \mu_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} P(z_i = j | X_i, \theta_k) x_i$$

$$* \nabla_{\Sigma_j} B(\theta, \theta_k) = -\sum_{i=1}^{n} P(z_i = j | X_i, \theta_k) \times (\Sigma_j^{-1})^2 \left[\frac{\Sigma_j}{2} - \frac{1}{2} (x_i - \mu_j)^T (x_i - \mu_j) \right] = 0$$

$$\Longrightarrow \Sigma_j^{k+1} = \frac{\sum_{i=1}^{n} P(z_i = j | X_i, \theta_k) (x_i - \mu_j)^T (x_i - \mu_j)}{\sum_{i=1}^{n} P(z_i = j | X_i, \theta_k)}$$

(derivation of the gradient:

$$B(\theta, \theta_k) = C + \sum_{i=1}^{n} \sum_{j=1}^{m} P(z_i = j | X_i, \theta_k) \left[-\frac{1}{2} (x_i - \mu_j)^T \Sigma_j^{-1} (x_i - \mu_j) + \frac{1}{2} \log(|\Sigma_j^{-1}|) \right]$$

because $|\Sigma_j|^{-1} = |\Sigma_j^{-1}|$

Then $\dfrac{\partial \log(|\Sigma_j^{-1}|)}{\partial \Sigma_j} = -\Sigma_j^T (\Sigma_j^{-1})^2 = -\Sigma_j (\Sigma_j^{-1})^2$

and $x^T A^{-1} x = tr(x x^T A^{-1})$

$$\frac{\partial x^T A^{-1} x}{\partial A} = \frac{\partial x^T A^{-1} x}{\partial A^{-1}} \frac{\partial A^{-1}}{\partial A} = x^T x (\Sigma_j^{-1})^2 \Big)$$

6) The number of degrees of freedom for m clusters is here $\underbrace{2m}_{\text{m means in 2D}} + \underbrace{3m}_{\text{m cov with cov being symmetrical}} + \underbrace{m-1}_{\text{the } \alpha_j\text{, the } -1 \text{ being there because } \sum_{j=1}^{m}\alpha_j = 1}$,

Exercise 3)

1) Step (ii) of the algorithm is the resolution of $\max_{\theta} \mathcal{L}(X|\theta)$

where $\mathcal{L}(X|\theta) = \sum_{i=1}^{n} \hat{\omega}_i \log\left(\sum_{j=1}^{M} \alpha_j f(X_i|\theta)\right)$

which strongly look like the likelihood for a gaussian mixture models. Solving this problem is very hard, and as it is from the same class of problem that the maximization of a gaussian mixture likelihood we can use the same tool ie the EM algorithm (the two problems are maximizations of functions with latent random variables).

The complete pseudo-likelihood becomes

$$\mathcal{L}(X, Z, \theta) = \sum_{i=1}^{n} \hat{\omega}_i \sum_{j=1}^{M} \log(\alpha_j f(X_i;\theta)) \mathbb{1}_{Z_i=j}$$

Taking the esperance,

$$B(\theta, \theta_k) = \mathbb{E}\left[\log \mathcal{L}(X, Z, \theta)\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{M} \hat{\omega}_i \,\mathbb{P}(Z_i = j | X_i, \theta_k) \log(\alpha_j f(X_i;\theta))$$

The updating step is then found in exactly the same way as for the gaussian mixture model, replacing $\mathbb{P}(Z_i = j \mid X_i, \theta_k)$ by $\hat{\omega}_i \, \mathbb{P}(Z_i = j \mid X_i, \theta_k)$,

ultimately,

$$\alpha_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_i \, \mathbb{P}(Z_i = j \mid X_i, \theta_k)$$

$$\mu_j^{k+1} = \frac{1}{n} \sum_{i=1}^{n} \hat{\omega}_i \, \mathbb{P}(Z_i = j \mid X_i, \theta_k) \, x_i$$

$$\Sigma_j^{k+1} = \frac{\sum_{i=1}^{n} \hat{\omega}_i \, \mathbb{P}(Z_i = j \mid X_i, \theta_k) (x_i - \mu_j)^T (x_i - \mu_j)}{\sum_{i=1}^{n} \hat{\omega}_i \, \mathbb{P}(Z_i = j \mid X_i, \theta_k)}$$