統碩一 106354012 曹立諭

EM for old faithful dataset

EM algorithm:

重複步驟直到參數收斂{

(E – step) For each i , set

$$\mathrm{Q}(\theta|\theta^{(t)}) = \mathrm{E_y}\left[\log p(x,y|\theta)\,\middle|x,\theta^{(t)}\right]$$

(M – step) Set

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)}) = \arg\max_{\theta} \sum_{i=1}^{n}\sum_{y_i=1}^{k}[\log\left(\alpha_{y_i}p(x_i|y_i,\theta)\right)p(y_i|x_i,\theta^{(t)})]$$

}

其中 (E-step) 為

$$\mathrm{Q}(\theta|\theta^{(t)}) \equiv \mathrm{E_y}\left[\log p(x,y|\theta)\,\middle|x,\theta^{(t)}\right]$$

$$= E[\sum_{i}\log\left(p(y_i|\theta)p(x_i|y_i,\theta)\right)|x,\theta^{(t)}]$$

$$= \sum_{i=1}^{n} E_y\left[\log(\alpha_{y_i}p(x_i|y_i,\theta))\middle|x,\theta^{(t)}\right]$$

$$= \sum_{i=1}^{n}\sum_{y_i=1}^{k}\left[\log\left(\alpha_{y_i}p(x_i|y_i,\theta)\right)p\left(y_i\middle|x_i,\theta^{(t)}\right)\right]$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{k}\left[\log\left(\alpha_j\phi\left(x_i\middle|\mu_i,\Sigma_j\right)\right)p\left(y_i=j\middle|x_i,\theta^{(t)}\right)\right]$$

Recall $\theta = \{\alpha_j, \mu_j, \Sigma_j; j = 1, \dots, k\}$

目的:找到使 Q 函數的最大值參數

(M-step)

我們從(E-step)得到概似函數，並將其最大概似函數的 MLE 計算出來，得到的結果如下

$$\alpha_j^{(t+1)} = \frac{\sum_{i=1}^{n} p\left(y_i = j \middle| x_i, \theta^{(t)}\right)}{n}$$

$$\mu_j^{(t+1)} = \frac{\sum_{i}^{n} x_i p\left(y_i = j \middle| x_i, \theta^{(t)}\right)}{\sum_{i=1}^{n} p\left(y_i = j \middle| x_i, \theta^{(t)}\right)}$$

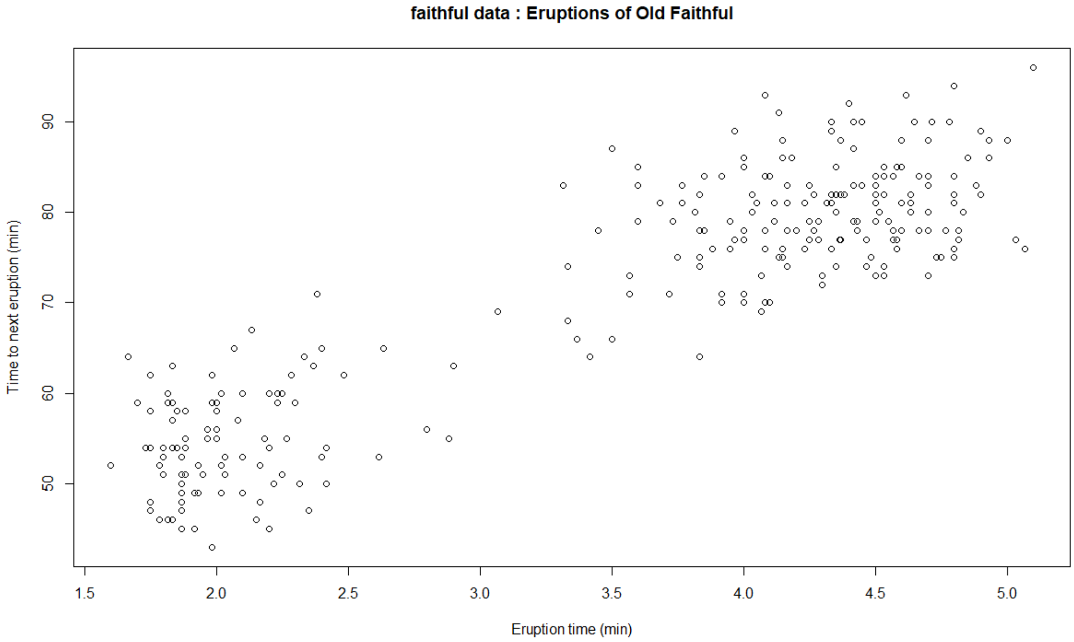
$$\Sigma_j^{(t+1)} = \frac{\sum_{i=1}^{n}\left(x_i - \mu_j^{(t)}\right)\left(x_i - \mu_j^{(t)}\right)^T p(y_i = j|x_i, \theta^{(t)})}{\sum_{i=1}^{n} p\left(y_i = j \middle| x_i, \theta^{(t)}\right)}$$

結論:

我們發現兩步驟相同的函數為 $p\left(y_i = j \middle| x_i, \theta^{(t)}\right)$，也就是說我們在撰寫 EM 演算法的時候，我們只要根據每次迭代得到新的參數，在給定原本的樣本下，把每一次隱藏變數的條件分配計算出來，並帶入我們(M-step)的結果算出新的參數，經過數次迭代後，就能找到收斂的參數。

Old Faithful dataset :

資料為美國黃石國家公園週期性噴發時間及等待時間資料(成對共 272 筆資料)



由上圖我們可以發生資料呈現為兩群的分佈，所以此題我們使用的分配為兩個混合的二元常態分配，並使用 EM 演算法來進行參數估計

$(\mathrm{x_i}, y_i) \sim a\ \ mix\ \ of\ \ N\left(\mu_j, \Sigma_\mathrm{j}\right)\ ,i = 1, \ldots, n\ \ ,j = 1,2\ \ ,\mu_j = \left(\mu_{j1}, \mu_{j2}\right)^\mathrm{T}$

我們先計算樣本中的平均數與變異數

	Eruption(min)	Waiting(min)
Means	3.487783	70.89706
Covariance matrix	$\begin{bmatrix} 1.302728 & 13.97781 \\ 13.97781 & 184.8233 \end{bmatrix}$	

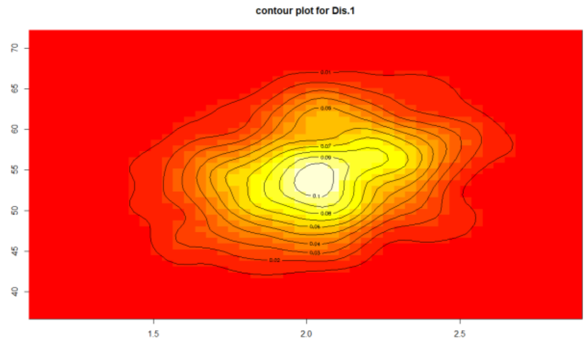
根據上圖，我們選取下列數值作為起始值

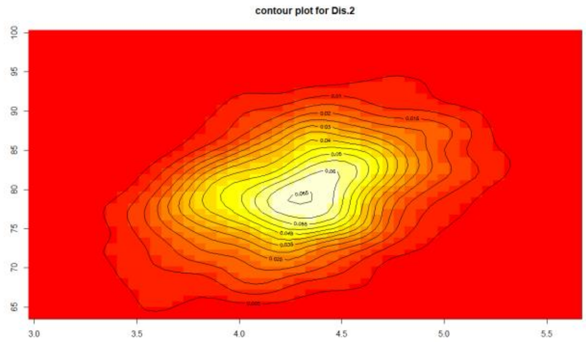
	Dis.1	Dis.2
Weights α_i	$\alpha_1 = 0.5$	$\alpha_2 = 0.5$
Means μ_j	$\mu_1 = (2,60)$	$\mu_2 = (4,80)$
Covariance matrix Σ_j	$\begin{bmatrix} 1 & 7 \\ 7 & 100 \end{bmatrix}$	$\begin{bmatrix} 2 & 20 \\ 20 & 200 \end{bmatrix}$

經由 EM 演算法迭代 30 次後，兩分配的估計參數如下:

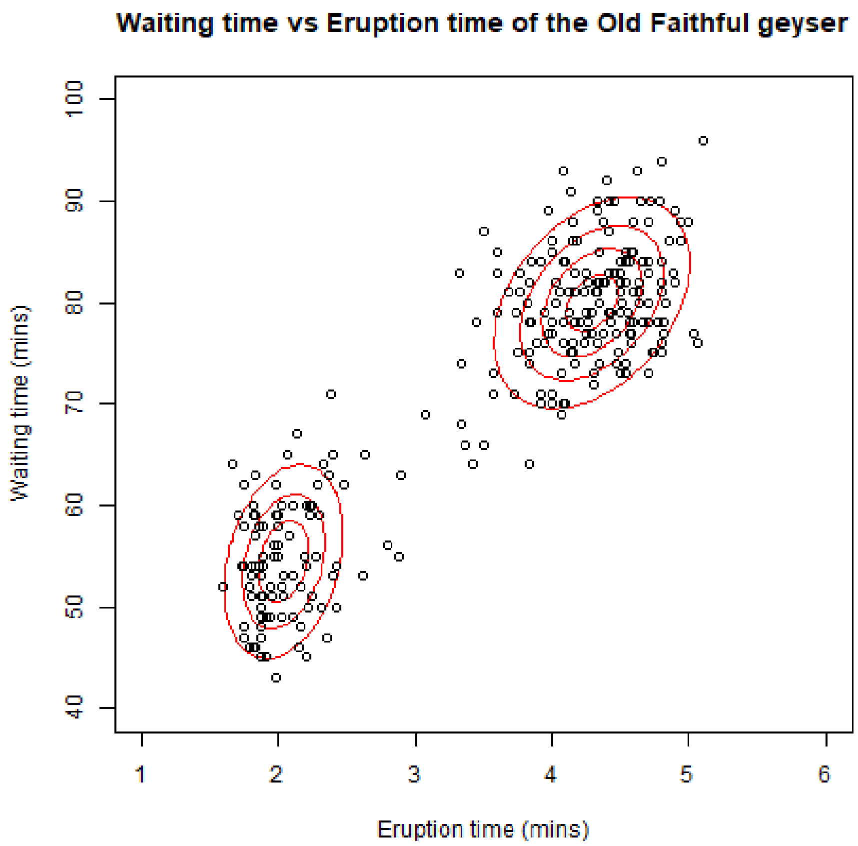
	Dis.1	Dis.2
Weights α_i	$\alpha_1 = 0.3558854$	$\alpha_2 = 0.6441146$
Means μ_j	$\mu_1 = (2.036421\ \ ,54.478880)$	$\mu_2 = (4.289688\ \ ,79.968413)$
Covariance matrix Σ_j	$\begin{bmatrix} 0.6991678 & 0.4400356 \\ 0.4400356 & 34.0510792 \end{bmatrix}$	$\begin{bmatrix} 0.1709121 & 0.9456217 \\ 0.9456217 & 36.2489521 \end{bmatrix}$

接著我們根據迭代後得到的參數，將其二元常態機率密度投影在二維平面上





我們將兩分配一同繪製在原始的圖表上



以上為從 EM 演算法估計出參數所繪製的結果

30 次迭代過程的動態 gif 檔 ： [網址]