GMM：高斯混合模型

高斯混合模型（Gaussian mixture model，簡稱GMM）是單一高斯機率密度函數的延伸，由於 GMM 能夠平滑地近似任意形狀的密度分佈，因此近年來常被用在語音與語者辨識，得到不錯的效果。

8 – 1. 單一高斯機率密度函數的參數估測法

假設我們有一組在高維空間（維度為 d）的點 $x_i, i=1\cdots n$，若這些點的分佈近似橢球狀，則我們可用高斯密度函數 $g(x_i \mu,\Sigma)$ 來描述產生這些點的機率密度函數：

$$g(x;\mu,\Sigma)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right]$$

其中 μ 代表此密度函數的中心點，Σ 則代表此密度函數的共變異矩陣（Covariance Matrix），這些參數決定了此密度函數的特性，如函數形狀的中心點、寬窄及走向等。

欲求得最佳的參數來描述所觀察到的資料點，可由最佳可能性估測法的概念來求得。在上述高斯密度函數的假設下，當 $x=x_i$ 時，其機率密度為 $g(x_i;\mu,\Sigma)$，若我們假設 $x_i, i=1\sim n$ 之間為互相獨立的事件，則發生 $X=\{x_1,x_2\cdots x_n\}$ 的機率密度為

$$p(X;\mu,\Sigma)=\prod_{i=1}^{n}g(x_i;\mu,\Sigma)$$

由於 X 是已經發生之事件，因此我們希望找出 μ,Σ 值，使得 $p(X;\mu,\Sigma)$ 能有最大值，此種估測參數 (μ,Σ) 的方法，即稱為**最佳可能性估測法**（MLE，Maximum Likelihood Estimation）

欲求得 $p(X;\mu,\Sigma)$ 的最大值，我們通常將之轉化為求下列 $J(\mu,\Sigma)$ 的最大值：

$$\begin{aligned}J(\mu,\Sigma)&=\ln p(X;\mu,\Sigma)\\&=\ln\left[\prod_{i=1}^{n}g(x_i;\mu,\Sigma)\right]\\&=\sum_{i=1}^{n}\ln g(x_i;\mu,\Sigma)\\&=\sum_{i=1}^{n}\left[-\frac{d}{2}\ln(2\pi)-\ln|\Sigma|-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\\&=-\frac{nd}{2}\ln(2\pi)-n\ln|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)\right]\end{aligned}$$

欲求最佳的 μ 值，直接求 $J(\mu,\Sigma)$ 對 μ 的微分即可：

$$\nabla_{\mu} J(\mu,\Sigma) = -\frac{1}{2}\sum_{i=1}^{n}\left[2\Sigma^{-1}(x_i-\mu)\right]$$
$$= -\Sigma^{-1}\left(\sum_{i=1}^{n} x_i - n\mu\right)$$

令上式等於零，我們就可以得到

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

欲求最佳的∑值，就不是那麼容易，需經過較繁雜的運算，在此我們僅列出結果：

$$\hat{\Sigma} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)\left(x_i-\hat{\mu}\right)^T$$

（對上式推導有興趣的同學，可以參考高等多變分析的相關教科書。）

8 – 2. 高斯混合密度函數的參數估測法

如果我們的資料$X=\{x_1,\cdots x_n\}$在 d 維空間中的分佈不是橢球狀，那麼就不適合以一個單一的高斯密度函數來描述這些資料點的機率密度函數。此時的變通方案，就是採用數個高斯函數的加權平均（Weighted Average）來表示。若以三個高斯函數來表示，則可表示成：

$$p(x) = \alpha_1 g(x;\mu_1,\Sigma_1) + \alpha_2 g(x;\mu_2,\Sigma_2) + \alpha_3 g(x;\mu_3,\Sigma_3)$$

此機率密度函數的參數為$(\alpha_1,\alpha_2,\alpha_3,\mu_1,\mu_2,\mu_3\,\Sigma_1,\Sigma_2,\Sigma_3)$，而且$\alpha_1,\alpha_2,\alpha_3$要滿足下列條件：

$$\alpha_1+\alpha_2+\alpha_3 = 1$$

以此種方式表示的機率密度函數，稱為「高斯混合密度函數」或是「高斯混合模型」（Gaussian Mixture Model），簡稱 GMM。

為簡化討論，我們通常假設各個高斯密度函數的共變異矩陣可以表示為：

$$\Sigma_j = \sigma_j^2 I = \sigma_j^2\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}, j=1,2,3$$

此時單一的高斯密度函數可表示如下：

$$g(x;\mu,\sigma^2) = (2\pi)^{-d/2}\sigma^{-d}\exp\left[-\frac{(x-\mu)^T(x-\mu)}{2\sigma^2}\right]$$

在上述方程式中，我們暫時省略了下標 j，以簡化方程式。若將上式對各個參數進行微分，可以得到下列等式：

$$\nabla_\mu g(x;\mu,\sigma^2) = g(x;\mu,\sigma^2)\left(-\frac{1}{2\sigma^2}\right)\nabla_\mu\left[(x-\mu)^T(x-\mu)\right]$$
$$= g(x;\mu,\sigma^2)\left(\frac{x-\mu}{\sigma^2}\right)$$
$$\nabla_\sigma g(x;\mu,\sigma^2) = (2\pi)^{-d/2}(-d)\sigma^{-d-1}e^{-\frac{(x-\mu)^T(x-\mu)}{2\sigma^2}} + (2\pi)^{-d/2}\sigma^{-d}e^{-\frac{(x-\mu)^T(x-\mu)}{2\sigma^2}}\left[\frac{(x-\mu)^T(x-\mu)}{\sigma^3}\right]$$
$$= g(x;\mu,\sigma^2)\left(\frac{(x-\mu)^T(x-\mu)}{\sigma^3} - \frac{d}{\sigma}\right)$$

上述這兩個等式，會在我們後面推導微分公式時，反覆被用到。

當共變異矩陣可以表示成一個常數和一個單位方陣的乘積時，前述的 $p(x)$ 可以簡化成：

$$p(x) = \alpha_1 g(x;\mu_1,\sigma_1^2) + \alpha_2 g(x;\mu_2,\sigma_2^2) + \alpha_3 g(x,\mu_3,\sigma_3^2)$$

此 $p(x)$ 的參數為 $\theta = [\alpha_1,\alpha_2,\alpha_3,\mu_1,\mu_2,\mu_3,\sigma_1^2,\sigma_2^2,\sigma_3^2]$，參數個數為 $1+1+1+d+d+d+1+1+1 = 6+3d$ 。

欲求得最佳的 θ 值，我們可依循最佳可能性估測法（MLE）原則，求出下列的最小值：

$$J(\theta) = \ln\left[\prod_{i=1}^{n} p(x_i)\right]$$
$$= \sum_{i=1}^{n} \ln p(x_i)$$
$$= \sum_{i=1}^{n} \ln\left[\alpha_1 g(x_i;\mu_1;\sigma_1^2) + \alpha_2 g(x_i;\mu_2,\sigma_2^2) + \alpha_3 g(x_i;\mu_3;\sigma_3^2)\right]$$

為簡化討論，我們引進另一個數學符號：

$$\beta_j(x) = \frac{\alpha_j g(x;\mu_j,\sigma_j^2)}{\alpha_1 g(x;\mu_1,\sigma_1^2) + \alpha_2 g(x;\mu_2,\sigma_2^2) + \alpha_3 g(x;\mu_3,\sigma_3^2)}$$

稱為事後機率（Post Probability），若用條件機率常用的表示方式，$\beta_j(x)$ 可寫成：

$$\beta_j(x) = p(j|x) = \frac{p(j\cap x)}{p(x)} = \frac{p(j)p(x|j)}{p(x)}$$
$$= \frac{p(j)p(x|j)}{p(1)p(x|1) + p(2)p(x|2) + p(3)p(x|3)}$$
$$= \frac{\alpha_j g(x;\mu_j,\sigma_j^2)}{\alpha_1 g(x;\mu_1,\sigma_1^2) + \alpha_2 g(x;\mu_2,\sigma_2^2) + \alpha_3 g(x;\mu_3,\sigma_3^2)}$$

因此$\beta_j(x)$可以看成是下列事件的機率：在觀測到亂數向量的值是x時，此向量是由第j個高斯密度函數所產生的。欲求$J(\theta)$的最小值，我們可以直接對μ_j及σ_j微分：

$$\nabla_{\mu_j} J(\theta) = \sum_{i=1}^{n} \frac{\alpha_j g\left(x;\mu_j,\sigma_j^2\right)}{\alpha_1 g\left(x_i;\mu_1,\sigma_1^2\right)+\alpha_2 g\left(x_i;\mu_2,\sigma_2^2\right)+\alpha_3 g\left(x_i;\mu_3,\sigma_3^2\right)} \frac{x_i-\mu_j}{\sigma_j^{\,2}}$$

$$= \sum_{i=1}^{n} \beta_j(x_i)\left(\frac{x_i-\mu_j}{\sigma_j^{\,2}}\right)$$

$$\nabla_{\sigma_j} J(\theta) = \sum_{i=1}^{n} \frac{\alpha_j g\left(x;\mu_j,\sigma_j^2\right)}{\alpha_1 g\left(x_i;\mu_1,\sigma_1^2\right)+\alpha_2 g\left(x_i;\mu_2,\sigma_2^2\right)+\alpha_3 g\left(x_i;\mu_3,\sigma_3^2\right)} \left[\frac{\left(x_i-\mu_j\right)^T\left(x_i-\mu_j\right)}{\sigma_j^3}-\frac{d}{\sigma_j}\right]$$

$$= \sum_{i=1}^{n} \beta_j(x_i)\left[\frac{\left(x_i-\mu_j\right)^T\left(x_i-\mu_j\right)}{\sigma_j^3}-\frac{d}{\sigma_j}\right]$$

令上兩式為零，即可得到：

$$\mu_j = \frac{\sum_{i=1}^{n} \beta_j(x_i) x_i}{\sum_{i=1}^{n} \beta_j(x_i)} \qquad (1)$$

$$\sigma_j^2 = \frac{1}{d} \frac{\sum_{i=1}^{n} \beta_j(x_i)\left(x_i-\mu_j\right)^T\left(x_i-\mu_j\right)}{\sum_{i=1}^{n} \beta_j(x_i)} \qquad (2)$$

此外，我們人必須求$J(\theta)$對α_j的微分，但因α_j仍必須滿足總和為 1 的條件，因此我們引進 Lagrange Multiplier ，並定義新的目標函數為：

$$J_{new} = J + \lambda\left(1-\alpha_1-\alpha_2-\alpha_3\right)$$

$$= \sum_{i=1}^{n} \ln\left[\alpha_1 g\left(x_i,\mu_1,\sigma_1^2\right)+\alpha_2 g\left(x_i,\mu_2,\sigma_2^2\right)+\alpha_3 g\left(x_i,\mu_3,\sigma_3^2\right)\right]+\lambda\left(1-\alpha_1-\alpha_2-\alpha_3\right)$$

$$\frac{\partial J_{new}}{\partial \alpha_j} = \sum_{i=1}^{n} \frac{g\left(x_i,\mu_j,\sigma_j^2\right)}{\alpha_1 g\left(x_i,\mu_1,\sigma_1^2\right)+\alpha_2 g\left(x_i,\mu_2,\sigma_2^2\right)+\alpha_3 g\left(x_i,\mu_3,\sigma_3^2\right)} - \lambda = 0$$

$$= \frac{1}{\alpha_j}\sum_{i=1}^{n} \beta_j(x_i) - \lambda = 0,\ j = 1,2,3$$

$$\Rightarrow \begin{cases} \alpha_1\lambda = \sum_{i=1}^{n} \beta_1(x_i) \\ \alpha_2\lambda = \sum_{i=1}^{n} \beta_2(x_i) \\ \alpha_3\lambda = \sum_{i=1}^{n} \beta_3(x_i) \end{cases}$$

將上三式相加：

$$(\alpha_1+\alpha_2+\alpha_3)\lambda=\sum_{i=1}^{n}[\beta_1(x_i)+\beta_2(x_i)+\beta_3(x_i)]$$

$$\lambda=\sum_{i=1}^{n}1=n$$

$$\Rightarrow \alpha_j=\frac{1}{n}\sum_{i=1}^{n}\ \beta_j(x_i),\ j=1,2,3 \qquad (3)$$

因此經由計算 $J(\theta)$ 的導式並令其為零，我們得到方程式 $(1),(2)$ 及 (3)，這三個方程式事實上代表了 $6+3d$ 個純量方程式，共含 $6+3d$ 個未知數，但須特別注意的是：$\beta_j(x)$ 仍是 $[\alpha_1,\alpha_2,\alpha_3,\mu_1,\mu_2,\mu_3,\sigma_1^2,\sigma_2^2,\sigma_3^2]$ 的函數，因此方程式 $(1),(2),(3)$ 是一組含 $6+3d$ 個未知數的非線性聯立方程式，很難用一般的方法去解，通常我們是以方程式 $(1),(2),(3)$ 為基礎來進行疊代法，流程如下：

1. 設定一個起始參數值 $\theta=[\alpha_1,\alpha_2,\alpha_3,\mu_1,\mu_2,\mu_3,\sigma_1^2,\sigma_2^2,\sigma_3^2]$。（我們可令 $\alpha_1=\alpha_2=\alpha_3=\frac{1}{3}$，並使用 K-means 的方式來計算群聚的中心點，以作為 μ_1、μ_2 和 μ_3 的起始參數值。）
2. 使用 θ 來計算 $\beta_1(x_i)$、$\beta_2(x_i)$ 及 $\beta_3(x_i)$，$i=1\sim n$
3. 計算新的 μ_j 值 ：

$$\tilde{\mu}_j=\frac{\sum_{i=1}^{n}\beta_j(x_i)x_i}{\sum_{i=1}^{n}\beta_j(x_i)}$$

4. 計算新的 σ_j 值：

$$\tilde{\sigma}_j^2=\frac{1}{d}\frac{\sum_{i=1}^{n}\beta_j(x_i)\left(x_i-\tilde{\mu}_j\right)^T\left(x_i-\tilde{\mu}_j\right)}{\sum_{i=1}^{n}\beta_j(x_i)}$$

5. 計算新的 α_j 值：

$$\tilde{\alpha}_j=\frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i)$$

6. 令 $\tilde{\theta}=[\tilde{\alpha}_1,\tilde{\alpha}_2,\tilde{\alpha}_3,\tilde{\mu}_1,\tilde{\mu}_2,\tilde{\mu}_2,\tilde{\sigma}_1^2,\tilde{\sigma}_2^2,\tilde{\sigma}_3^2]$ 若 $\left\|\theta-\tilde{\theta}\right\|$ 小於某一個極小的容忍值，則停止。否則令 $\theta=\tilde{\theta}$ 並跳回步驟 2。

上述疊代方法一定會讓$J(\theta)$逐步遞增，並收斂至一個局部最大值（Local Maximum），但我們無法證明此局部最大值是否就是全域最大值（Global Maximum）。有關這些方程式的另一種推導方法，以及這些方程式能夠讓$J(\theta)$逐步遞增的證明，詳見下節說明。

8 – 3. 求取 GMM 參數的另一種方法

在本節中，我們使用另一種來導出求取 GMM 參數的疊代公式。此方法所得到的疊代公式與前一節的公式完全相同，但本節之方法可證明此疊代公式可以逐次提高$J(\theta)$的值。

首先我們說明一個重要的不等式。由於對數函數$f(x)=\ln(x)$是一個凸函數（Convex Function），滿足下列不等式：

$$\ln[\lambda x_1+(1-\lambda)x_2]\geq\lambda\ln(x_1)+(1-\lambda)\ln(x_2)$$

推廣上式可得「簡森不等式」（Jensen’s Inequality）：

$$\ln\left(\sum_{i=1}^{n}\lambda_i x_i\right)\geq\sum_{i=1}^{n}\lambda_i\ln(x_i)$$

其中λ_i必須滿足$\sum_{i=1}^{n}\lambda_i=1$。

假設我們現有的參數是$\hat{\theta}$，我們希望找出新的θ值，使得$J(\theta)>J(\hat{\theta})$。以$m=3$為例，$J(\theta)$可以表示成：

$$J(\theta)=\sum_{i=1}^{n}\ln\left[\alpha_1 g(x_i,\mu_1,\sigma_1^2)+\alpha_2 g(x_i,\mu_2,\sigma_2^2)+\alpha_3 g(x_i,\mu_3,\sigma_3^2)\right]$$

因此

$$\begin{aligned}
J(\theta)-J(\hat{\theta}) &= \sum_{i=1}^{n}\ln\left[\frac{\alpha_1 g(x_i,\mu_1,\sigma_1^2)+\alpha_2 g(x_i,\mu_2,\sigma_2^2)+\alpha_3 g(x_i,\mu_3,\sigma_3^2)}{\hat{\alpha}_1 g(x_i,\hat{\mu}_1,\hat{\sigma}_1^2)+\hat{\alpha}_2 g(x_i,\hat{\mu}_2,\hat{\sigma}_2^2)+\hat{\alpha}_3 g(x_i,\hat{\mu}_3,\hat{\sigma}_3^2)}\right]\\
&= \sum_{i=1}^{n}\ln\left[\frac{\alpha_1 g(x_i,\mu_1,\sigma_1^2)}{D(\hat{\theta})}\frac{\beta_1(x_i)}{\beta_1(x_i)}+\frac{\alpha_2 g(x_i,\mu_2,\sigma_2^2)}{D(\hat{\theta})}\frac{\beta_2(x_i)}{\beta_2(x_i)}+\frac{\alpha_3 g(x_i,\mu_3,\sigma_3^2)}{D(\hat{\theta})}\frac{\beta_3(x_i)}{\beta_3(x_i)}\right]\\
&\geq \sum_{i=1}^{n}\left[\beta_1(x_i)\ln\frac{\alpha_1 g(x_i,\mu_1,\sigma_1^2)}{D(\hat{\theta})\beta_1(x_i)}+\beta_2(x_i)\ln\frac{\alpha_2 g(x_i,\mu_2,\sigma_2^2)}{D(\hat{\theta})\beta_2(x_i)}+\beta_3(x_i)\ln\frac{\alpha_3 g(x_i,\mu_3,\sigma_3^2)}{D(\hat{\theta})\beta_3(x_i)}\right]\\
&= Q(\theta)
\end{aligned}$$

在前面的推導中，$\beta_j(x_i)$的計算是根據$\hat{\theta}$，而且

$$D(\hat{\theta})=\hat{\alpha}_1 g(x_i,\hat{\mu}_1,\hat{\sigma}_1^2)+\hat{\alpha}_2 g(x_i,\hat{\mu}_2,\hat{\sigma}_2^2)+\hat{\alpha}_3 g(x_i,\hat{\mu}_3,\hat{\sigma}_3^2)$$

因為$\sum_{j=1}^{3}\beta_j(x_i)=1$，因此我們可套用簡森不等式而得到$J(\theta)-J(\hat{\theta})\geq Q(\theta)$。換句話說，只要$Q(\theta)>0$，那麼$J(\theta)$就會大於$J(\hat{\theta})$，但通常我們希望$J(\theta)$是越大越好，因

此最直覺的方法是直接求得使$Q(\theta)$為最大的θ值，那麼$J(\theta)$就會跟著變大，見圖1。由於$Q(\theta)$是θ的函數，我們可以把和θ不相關的部分併入常數項，如下：

$$Q(\theta)=\sum_{i=1}^{n}\left\{\beta_1(x_i)\ln\left[\alpha_1 g\left(x_i,\mu_1,\sigma_1^2\right)\right]+\beta_2(x_i)\ln\left[\alpha_2 g\left(x_i,\mu_2,\sigma_2^2\right)\right]+\beta_3(x_i)\ln\left[\alpha_3 g\left(x_i,\mu_3,\sigma_3^2\right)\right]\right\}+c_1$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left[\ln\alpha_j+\ln g\left(x_i,\mu_j,\sigma_j^2\right)\right]+c_1$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left\{\ln\alpha_j+\ln\left[\frac{1}{\left(2\pi\sigma_j^2\right)^{d/2}}\exp\left(-\frac{\left(x_j-\mu_j\right)^T\left(x_j-\mu_j\right)}{2\sigma_j^2}\right)\right]\right\}+c_1$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left[\ln\alpha_j-d\ln\sigma_j-\frac{\left(x_i-\mu_j\right)^T\left(x_i-\mu_j\right)}{2\sigma_j^2}\right]+c_2$$

$$\nabla_{\mu_j}Q=0\Rightarrow\mu_j=\frac{\sum_{i=1}^{n}\beta_j(x_i)x_i}{\sum_{i=1}^{n}\beta_j(x_i)}$$

$$\nabla_{\sigma_j}Q=0\Rightarrow\sigma_j^2=\frac{1}{d}\frac{\sum_{i=1}^{n}\beta_j(x_i)\left(x_i-\mu_j\right)^T\left(x_i-\mu_j\right)}{\sum_{i=1}^{n}\beta_j(x_i)}$$

欲求最佳之α_j值，需引入Lagrange multiplier：

$$Q_{new}=Q+\lambda(\alpha_1+\alpha_2+\alpha_3-1)$$

$$=\sum_{i=1}^{n}\sum_{j=1}^{3}\beta_j(x_i)\left[\ln\alpha_j-d\ln\sigma_j-\frac{\left(x_i-\mu_j\right)^T\left(x_i-\mu_j\right)}{2\sigma_j^2}\right]+\lambda(\alpha_1+\alpha_2+\alpha_3-1)$$

分別對α_j微分，可得

$$\nabla_{\alpha_1}Q_{new}=0\Rightarrow\alpha_1\lambda=-\sum_{i=1}^{n}\beta_1(x_i)$$

$$\nabla_{\alpha_2}Q_{new}=0\Rightarrow\alpha_2\lambda=-\sum_{i=1}^{n}\beta_2(x_i)$$

$$\nabla_{\alpha_3}Q_{new}=0\Rightarrow\alpha_3\lambda=-\sum_{i=1}^{n}\beta_3(x_i)$$

將上三式相加，可得：

$$(\alpha_1+\alpha_2+\alpha_3)\lambda=-\sum_{i=1}^{n}\left[\beta_1(x_i)+\beta_2(x_i)+\beta_3(x_i)\right]$$

$$\lambda=-\sum_{i=1}^{n}1=-n$$

因此

$$\alpha_j=\frac{1}{n}\sum_{i=1}^{n}\beta_j(x_i),\ j=1,2,3$$

因此我們最後的結果可整理如下：

$$\begin{cases} \mu_j = \dfrac{\sum_{i=1}^{n} \beta_j(x_i) x_i}{\sum_{i=1}^{n} \beta_j(x_i)} \\ \sigma_j^2 = \dfrac{1}{d} \dfrac{\sum_{i=1}^{n} \beta_j(x_i)(x_i - \mu_j)^T (x_i - \mu_j)}{\sum_{i=1}^{n} \beta_j(x_i)} \\ \alpha_j = \dfrac{1}{n} \sum_{i=1}^{n} \beta_j(x_i) \end{cases}$$

其中 $\beta_j(x_i)$ 的計算，是根據 $\hat{\theta}$。因此由上述方法得到的結果，和前一節的結果是完全一致的。有關於 $J(\theta) - J(\hat{\theta}) \geq Q(\theta)$ 這部分的圖解，可見下圖。

圖一.