模式识别　　第三次作业

宁雪妃　2016310503

1. EM and Gradient Descent

对于 gradient descent，写出 $\nabla_{\mu_k} l$ ~~和 $\nabla_{\sigma_k} l$~~ 如下：

$$\nabla_{\mu_k} l = \sum_{i=1}^{n} \frac{\pi_k N(x_i|\mu_k, \sigma_k^2 I)}{\sum_{k'=1}^{K} \pi_{k'} N(x_i|\mu_{k'}, \sigma_{k'}^2 I)} \cdot \frac{\partial N(x_i|\mu_k, \sigma_k^2 I)}{\partial \mu_k}$$

$$= \sum_{i=1}^{n} \frac{\pi_k N(x_i|\mu_k, \sigma_k^2 I)}{\sum_{k'=1}^{K} \pi_{k'} N(x_i|\mu_{k'}, \sigma_{k'}^2 I)} \frac{x_i - \mu_k}{\sigma_k^2}$$

对于EM：

$$\hat{z}_{ik}^{(t+0.5)} = P(x_i \in cluster_k | x_i, \{(\mu_j^{(t)}, (\sigma_j^2)^{(t)})\}_{j=1}^{K})$$

$$= \frac{P(x_i \in cluster_k, x_i)}{\sum_{k'=1}^{K} P(x_i \in cluster_{k'}, x_i)} = \frac{\pi(k) N(x_i|\mu_k, \sigma_k^2 I)}{\sum_{k'=1}^{K} \pi(k) N(x_i|\mu_k, \sigma_k^2 I)}$$

$$Q = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik}^{(t+0.5)} \left(\log N(x_i|\mu_k, (\sigma_k^2)^{(t)} I) + \log \pi_k\right)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik}^{(t+0.5)} \left(-\frac{(x_i-\mu_k)^T (x_i-\mu_k)}{2\sigma_k^2} + Const(\mu_k)\right)$$

明显有 Q 关于 μ_k 为二次函数，其最值点唯一，且在极值点取得，有

$$\frac{\partial Q}{\partial \mu_k} = \sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)} \frac{(x_i - \mu_k)}{\sigma_k^2} = 0$$ 为极值点 $\Rightarrow \mu_{k(em)}^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)} x_i}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)}}$

对于 gradient descent：

$$\mu_{k(gd)}^{(t+1)} = \mu_{k(gd)}^{(t)} + \frac{\eta_k^{(t)}}{\sigma_k^2} \sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)} (x_i - \mu_k^{(t)}) = \mu_{k(gd)}^{(t)} \left(1 - \frac{\eta_k^{(t)}}{\sigma_k^2} \sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)}\right) + \frac{\eta_k^{(t)}}{\sigma_k^2} \sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)} x_i$$

取 $\eta_k^{(t)} = \sigma_k^2 \Big/ \sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)}$

$$\Rightarrow \mu_{k(gd)}^{(t+1)} = \frac{\sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)} x_i}{\sum_{i=1}^{n} \hat{z}_{ik}^{(t+0.5)}} = \mu_{k(em)}^{(t+1)}$$

令 $S_k=\sigma_k^2$. 有 $\nabla_S l = \sum_{i=1}^{n} \frac{\pi(k)\, N(x_i|\mu_k^{(t+1)}, S_k^{(t)} I)}{\sum_{k'=1}^{K} \pi_{k'}\, N(x_i|\mu_{k'}^{(t+1)}, S_{k'}^{(t)} I)} \left(\frac{\|x_i-\mu_k^{(t+1)}\|^2}{2(S_k^{(t)})^2} - \frac{P}{2 S_k^{(t)}} \right)$

对于EM:

$$Q=\sum_{i=1}^{n}\sum_{k=1}^{K}\hat{z}_{ik}^{(t+1)}\left(-\frac{\|x_i-\mu_k\|^2}{2S_k}-\frac{P}{2}\ln S_k\right)$$

$$\frac{\partial Q}{\partial S_k}=\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\|x_i-\mu_k\|^2}{2S_k^2}-\frac{P\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}}{2S_k}$$

当 $S_{k*}=\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\|x_i-\mu_k\|^2}{P\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}}$ 时 $\frac{\partial Q}{\partial S_k}=0$. 当 $\begin{cases} S_k>S_{k*} \text{ 时 } \frac{\partial Q}{\partial S_k}<0 \\ S_k<S_{k*} \text{ 时 } \frac{\partial Q}{\partial S_k}>0 \end{cases}$（注意满足 $S_k=\sigma_k^2>0$）

$\therefore S_{k*}$ 为全局最优值.

即 $\left(\sigma_{k[em]}^{(t+1)}\right)^2=\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\|x_i-\mu_k^{(t+1)}\|^2}{P\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}}$. 其中P为x/μ的维度

$$[\sigma_k^2]_{[gd]}^{(t+1)}=[\sigma_k^2]^{(t)}+S_k^{(t)}\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\left(\frac{\|x_i-\mu_k^{(t+1)}\|^2}{2(\sigma_k^{(t)})^4}-\frac{P}{2(\sigma_k^{(t)})^2}\right)$$

取 $S_k^{(t)}=\frac{2(\sigma_k^{(t)})^4}{P\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}}$ $\Rightarrow (\sigma_k^2)_{[gd]}^{(t+1)}=(\sigma_k^2)^{(t)}\left[1-\frac{P\left(\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\right)S_k^{(t)}}{2(\sigma_k^{(t)})^4}\right]$

$$+\,S_k^{(t)}\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\|x_i-\mu_k\|^2}{2(\sigma_k^{(t)})^4}$$

$$=\frac{\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}\|x_i-\mu_k\|^2}{P\sum_{i=1}^{n}\hat{z}_{ik}^{(t+1)}}=\left(\sigma_{k[em]}^{(t+1)}\right)^2$$

得证□

2. Mixture of multinomial variables

1. $l(\pi,\mu) = \log \prod_{i=1}^{M} \sum_{k=1}^{K} \pi_k \prod_{n=1}^{N} \mu_{kn}^{x_n^{(i)}} = \sum_{i=1}^{m} \log\left(\sum_{k=1}^{K} \pi_k \prod_{n=1}^{N} \mu_{kn}^{x_n^{(i)}} \right)$

2. $l(\pi,\mu) = \sum_{i=1}^{M} \log(P(x^{(i)}|\pi,\mu)) = \sum_{i=1}^{M} \log\left(\sum_{k=1}^{K} \cancel{I(x^{(i)}\text{由 cluster } k \text{ 生成})}\, P(x^{(i)}, x^{(i)}\text{由 cluster } k \text{ 生成}) \right)$

$\geq \sum_{i=1}^{M} \sum_{k=1}^{K} q(x^{(i)}\text{由 cluster } k \text{ 生成}) \log\left(\frac{P(x^{(i)}, x^{(i)}\text{由 cluster } k \text{ 生成})}{q(x^{(i)}\text{由 cluster } k \text{ 生成})} \right)$

记 $q^*(x^{(i)}\text{由 cluster } k \text{ 生成}) = \hat{z}_{ik}$ 为E-step q的最优解. 且由凹函数取更换 $f(E(X)) = E(f(X))$ 一定说明 X为单点的性质知:

$\hat{z}_{ik} = q^*(x^{(i)}\text{由 cluster } k \text{ 生成}) = P(x^{(i)}\text{由 cluster } k \text{ 生成} / x^{(i)}, \{\mu_{jn}\}_{\substack{j=1\cdots K\\ n=1\cdots N}})$

$= \frac{\pi_k(t) \prod_{n=1}^{N} \mu_{kn}(t)^{x_n^{(i)}}}{\sum_{k'=1}^{K} \pi_{k'}(t) \prod_{n=1}^{N} \mu_{k'n}(t)^{x_n^{(i)}}}$

M-step:

$Q(\theta(t), \theta(t+1)) = \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \left[\log(\pi_k^{(t+1)}) + \log\left(\prod_{n=1}^{N} \mu_{kn}(t+1)^{x_n^{(i)}}\right) \right] + Const(\theta(t+1))$

(对于 $\theta(t+1)$为 const.)

$= \sum_{k=1}^{K} \log(\pi_k(t+1)) \sum_{i=1}^{M} \hat{z}_{ik} + \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \sum_{n=1}^{N} x_n^{(i)} \ln(\mu_{kn}(t+1)) + Const(Q\ldots)$

4. $\frac{\partial \text{Lagranian}(Q)}{\partial \pi_k(t+1)} = \frac{\partial\left(\sum_{k=1}^{K} \log(\pi_k(t+1)) \sum_{i=1}^{M} \hat{z}_{ik} - \lambda\left(\sum_{k=1}^{K} \pi_k(t+1) - 1\right) \right)}{\partial \pi_k(t+1)}$

$= \frac{\sum_{i=1}^{M} \hat{z}_{ik}}{\pi_k(t+1)} - \lambda = 0$ 即 $\pi_k(t+1) = \frac{1}{\lambda} \sum_{i=1}^{M} \hat{z}_{ik}$

$\Rightarrow 1 = \sum_{k=1}^{K} \pi_k(t+1) = \frac{1}{\lambda} \sum_{k=1}^{K} \sum_{i=1}^{M} \hat{z}_{ik} \Rightarrow \lambda = \sum_{k=1}^{K} \sum_{i=1}^{M} \hat{z}_{ik}$

$\therefore \pi_k(t+1) = \frac{\sum_{i=1}^{M} \hat{z}_{ik}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \hat{z}_{ik}}$.

$\frac{\partial \text{Lagranian}(Q)}{\partial \mu_{kn}(t+1)} = \frac{\partial\left(\sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \sum_{n=1}^{N} x_n^{(i)} \ln(\mu_{kn}(t+1)) - \sum_{k=1}^{K} \lambda \left(\sum_{n=1}^{N} \mu_{kn} - 1\right) \right)}{\partial \mu_{kn}(t+1)}$

$$= \frac{\sum_{i=1}^{M} \hat{z}_{ik} x_n^{(i)}}{\mu_{kn}(t+1)} - \lambda_k = 0 \quad \Rightarrow \mu_{kn}(t+1) = \frac{1}{\lambda_k} \sum_{i=1}^{M} \hat{z}_{ik} x_n^{(i)}$$

$$1 = \sum_{n=1}^{N} \mu_{kn}(t+1) = \frac{1}{\lambda_k} \sum_{n=1}^{N} \sum_{i=1}^{M} \hat{z}_{ik} x_n^{(i)} = \frac{1}{\lambda_k} \sum_{i=1}^{M} \hat{z}_{ik} \sum_{n=1}^{N} x_n^{(i)} = \frac{1}{\lambda_k} \sum_{i=1}^{M} \hat{z}_{ik}.$$

$$\Rightarrow \lambda_k = \sum_{i=1}^{M} \hat{z}_{ik} \quad \Rightarrow \mu_{kn}(t+1) = \frac{\sum_{i=1}^{M} \hat{z}_{ik} x_n^{(i)}}{\sum_{i=1}^{M} \hat{z}_{ik}} \quad \begin{cases} k = 1, \cdots, K \\ n = 1, \cdots, N \end{cases}$$

3. ~~有~~ 在前面用Jensen不等式可知 代入 $\forall\, q(x^{(i)}$由 cluster k 生成). 都有

$$L(\pi, \mu) \geq D(\pi, \mu, q) \qquad D\text{为下界函数}.$$

Q为将E step的最优 q^* 代入D得到的函数. $\therefore$ 一定有 $L(\pi,\mu) \geq Q(\pi,\mu)$

设M-step 找到 Q的最大值 π^*, μ^*, 有

$$L(\pi^*, \mu^*) \geq Q(\pi^*, \mu^*) \geq Q(\pi^{old}, \mu^{old}) = L(\pi^{old}, \mu^{old})$$

(不好意思, 写到这里发现老师的Q函数定义和我不一样… 差了一个 $f(\pi^{old}, \mu^{old})$ 为 $\sum_{i=1}^{M} H(q^*(i))$) 从这里往下改变Q定义

$$Q = \sum_{i=1}^{M} \sum_{k=1}^{K} \hat{z}_{ik} \log\left(\pi_k \prod_{n=1}^{N} \mu_{kn}^{x_n^{(i)}}(t+1)\right) = D(\pi, \mu, q^*) + \sum_{i=1}^{M} \sum_{k=1}^{K} q^* \log q^*$$

$\therefore$ 应为 $L(\pi^*, \mu^*) \geq D(\pi, \mu, q^*) = Q - \sum_{i=1}^{M} \sum_{k=1}^{K} q^*(i) \log q^*(i)$

$$\therefore \quad L(\pi,\mu) - L(\pi^{old}, \mu^{old}) \geq Q - \sum_{i=1}^{M} \sum_{k=1}^{K} q^*(i) \log q^*(i) - L(\pi^{old}, \mu^{old})$$

其中 $L(\pi^{old}, \mu^{old})$ 与 $q^*(i)$ 都只与 π^{old}, μ^{old} 有关.

$$\therefore \text{令}\ f(\pi^{old}, \mu^{old}) = -\left(\sum_{i=1}^{M} \sum_{k=1}^{K} q^*(i) \log q^*(i) + L(\pi^{old}, \mu^{old})\right)$$

$$\Rightarrow L(\pi,\mu) - L(\pi^{old}, \mu^{old}) \geq Q + f(\pi^{old}, \mu^{old}) \qquad \text{得证} \ \square$$

3. ① E-step.

a. $X = Y + Z + \varepsilon$. 其中 $Y \sim N(\mu_p, \sigma_p^2)$, $Z \sim N(v_r, \tau_r^2)$, $\varepsilon \sim N(0, \sigma^2)$

$E(X) = E(Y+Z+\varepsilon) = \mu_p + v_r$ $\qquad Y \perp Z \perp \varepsilon$

$$E(X^2) = E[(Y+Z+\varepsilon)^2] = E[Y^2] + E[Z^2] + E[\varepsilon^2] + 2E[Y]E[Z]$$

$$= Var(Y) + E^2(Y) + Var(Z) + E^2(Z) + Var(\varepsilon) + E^2(\varepsilon) + 2E[Y]E[Z]$$

$$= \sigma_p^2 + \mu_p^2 + \tau_r^2 + v_r^2 + \sigma^2 + 2\mu_p v_r$$

$$Var(X) = E(X^2) - E^2(X) = \sigma_p^2 + \mu_p^2 + \tau_r^2 + v_r^2 + \sigma^2 - \mu_p^2 - v_r^2 - 2\mu_p v_r + 2\mu_p v_r$$

$$= \sigma_p^2 + \tau_r^2 + \sigma^2 \cancel{- 2\mu_p v_r}$$

b. E-step: 推导隐变量 $y^{(pr)}$ 与 $z^{(pr)}$ 的条件概率

$$p(y^{(pr)}, z^{(pr)} \mid x^{(pr)}) = \frac{p(x^{(pr)} \mid y^{(pr)}, z^{(pr)})\, p(y^{(pr)})\, p(z^{(pr)})}{p(x^{(pr)})}$$

$$\begin{bmatrix} Y \\ Z \\ X \end{bmatrix} = \begin{bmatrix} Y \\ Z \\ Y+Z+\varepsilon \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 1 & \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} Y \\ Z \\ \varepsilon \end{bmatrix}$$ 用条件高斯公式有

$$\begin{bmatrix} Y \\ Z \end{bmatrix} \Big| X \sim N\left(E\begin{bmatrix} Y \\ Z \end{bmatrix} + \Sigma_{12}\Sigma_{22}^{-1}(X - E(X)),\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

其中 $E\begin{bmatrix} Y \\ Z \end{bmatrix} = \begin{bmatrix} \mu_p \\ v_r \end{bmatrix}$ $\quad E[X] = \mu_p + v_r$.

$$\Sigma = A\,\Sigma_0\left(\begin{bmatrix} Y \\ Z \\ \varepsilon \end{bmatrix}\right)A^T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \sigma_p^2 & & \\ & \tau_r^2 & \\ & & \sigma^2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_p^2 & 0 & \sigma_p^2 \\ 0 & \tau_r^2 & \tau_r^2 \\ \sigma_p^2 & \tau_r^2 & \sigma_p^2 + \tau_r^2 + \sigma^2 \end{bmatrix}$$

$\Sigma_{11} = \begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix}$ $\quad \Sigma_{22} = \sigma_p^2 + \tau_r^2 + \sigma^2$ $\quad \Sigma_{12} = [\sigma_p^2, \tau_r^2]^T$

$\Sigma_{21} = [\sigma_p^2, \tau_r^2]$

$$\therefore Q_{pr}(y^{(pr)}, z^{(pr)}) = N\left(\begin{bmatrix} \mu_p \\ v_r \end{bmatrix} + \frac{(x^{(pr)} - (\mu_p + v_r))}{\sigma_p^2 + \tau_r^2 + \sigma^2}\begin{bmatrix} \sigma_p^2 \\ \tau_r^2 \end{bmatrix},\right.$$

$$\left.\begin{bmatrix} \sigma_p^2 & 0 \\ 0 & \tau_r^2 \end{bmatrix} - \frac{1}{\sigma_p^2 + \tau_r^2 + \sigma^2}\begin{bmatrix} \sigma_p^4 & \tau_r^2\sigma_p^2 \\ \tau_r^2\sigma_p^2 & \tau_r^4 \end{bmatrix}\right)$$

② M-step

$$Q = \sum_{p=1}^{P}\sum_{r=1}^{R}\int_{y,z} Q_{pr}\left(\log p(x^{(pr)}|y^{(pr)}, z^{(pr)}) + \log p(y^{(pr)}|\mu_p, \sigma_p^2) + \log p(z^{(pr)}|v_r, \tau_r^2)\right) dy\, dz$$

$$\frac{\partial Q}{\partial \mu_p} = \sum_{r=1}^{R}\int_{y^{(pr)}, z} Q_{pr} \frac{y^{(pr)}-\mu_p}{\sigma_p^2} dy^{(pr)} dz^{(pr)} = 0$$

$$= \sum_{r=1}^{R}\int_{y^{(pr)}} Q_{pr}(y^{(pr)}) y^{(pr)} dy - R\mu_p = 0 \Rightarrow \mu_p = \frac{1}{R}\sum_{r=1}^{R}\int_y Q_{pr}(y) y\, dy$$

$$= \frac{1}{R}\sum_{r=1}^{R} E_{Q_{pr}}[y|x]$$

$$E[y|x^{(pr)}] = \mu_p + \frac{x^{(pr)} - \mu_p + v_r}{\sigma_p^2 + \tau_r^2 + \sigma^2}\sigma_p^2$$

$$\therefore \mu_p^{(t+1)} = \frac{1}{R}\sum_{r=1}^{R}\left(\mu_p^{(t)} + \frac{x^{(pr)} - (\mu_p^{(t)} + v_r^{(t)})}{\sigma_p^{2(t)} + \tau_r^{2(t)} + \sigma^2}\sigma_p^{2(t)}\right)$$

对于 v_r 类似有

$$\frac{\partial Q}{\partial v_r} = \sum_{p=1}^{P}\int_z Q_{pr}(z|x) z\, dz - P v_r = 0 \Rightarrow v_r = \frac{1}{P}\sum_{p=1}^{P} E[z|x^{(pr)}]$$

$$= \frac{1}{P}\sum_{p=1}^{P}\left(v_r^{(t)} + \frac{x^{(pr)} - (\mu_p^{(t)} + v_r^{(t)})}{\sigma_p^{2(t)} + \tau_r^{2(t)} + \sigma^2}\tau_r^2\right)$$

$$\frac{\partial Q}{\partial (\sigma_p^2)} = \sum_{r=1}^{R}\int_y Q_{pr}(y|x)\left(\frac{\|y-\mu_p\|^2}{2\sigma_p^4} - \frac{1}{2\sigma_p^2}\right) dy = 0$$

$$\Rightarrow R\sigma_p^{2(t+1)} = \sum_{r=1}^{R} E_{Q_{pr}(y|x)}[\|y - \mu_p^{(t+1)}\|^2 | x^{(pr)}] \Rightarrow \cancel{\sigma_p^{2(t+1)} = \frac{1}{R}\sum_{r=1}^{R}\frac{\sigma_p^2\tau_r^2 + \sigma_p^2\sigma^2}{\sigma_p^2 + \tau_r^2 + \sigma^2}}$$

~~有错误~~

~~同理~~ $\cancel{\tau_r^{2(t+1)} = \frac{1}{P}\sum_{r=1}^{R}\frac{\tau_r^{2(t)}\sigma_p^{2(t)} + \tau_r^{2(t)}\sigma^2}{\sigma_p^2 + \tau_r^2 + \sigma^2}}$

$$R\sigma_p^{2(t+1)} = \sum_{r=1}^{R} \int_y Q_{pr}(y|x)\, \|y-\mu_p^{(t)}+\mu_p^{(t)}-\mu_p^{(t+1)}\|^2\, dy$$

$$= \sum_{r=1}^{R} \int_y Q_{pr}(y|x)\left[\|y-\mu_p^{(t)}\|^2 + 2(\mu_p^{(t)}-\mu_p^{(t+1)})^T(y-\mu_p^{(t)}) + \|\mu_p^{(t)}-\mu_p^{(t+1)}\|^2\right] dy$$

$$= \sum_{r=1}^{R} Var(Y|X)^{(t)} + 0 + \|\mu_p^{(t)}-\mu_p^{(t+1)}\|^2.$$

$$= \sum_{r=1}^{R}\left(\sigma_p^{2(t)} - \frac{\sigma_p^{4(t)}}{\sigma_p^{2(t)}+\tau_r^{2(t)}+\sigma^2}\right) + R\|\mu_p^{(t)}-\mu_p^{(t+1)}\|^2$$

同理 $P\tau_r^{2(t+1)} = \sum_{p=1}^{P}\left(\tau_r^{2(t)} - \frac{\tau_r^{4(t)}}{\sigma_p^{2(t)}+\tau_r^{2(t)}+\sigma^2}\right) + P\|\nu_r^{(t)}-\nu_r^{(t+1)}\|^2$

Programming 1

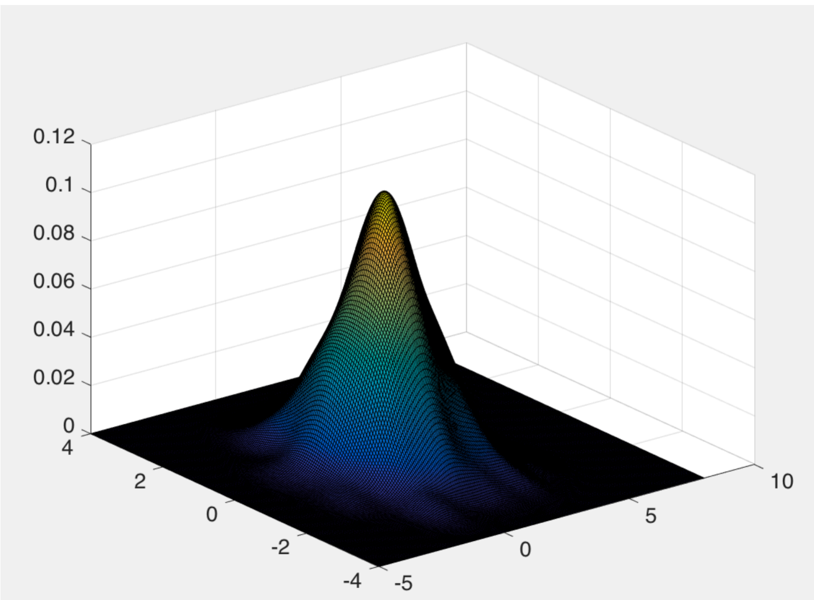
1. 理想 bayesian 分类器的 error rate 用 Riemann 积分方法计算得到：
   err_rate = 0.1988

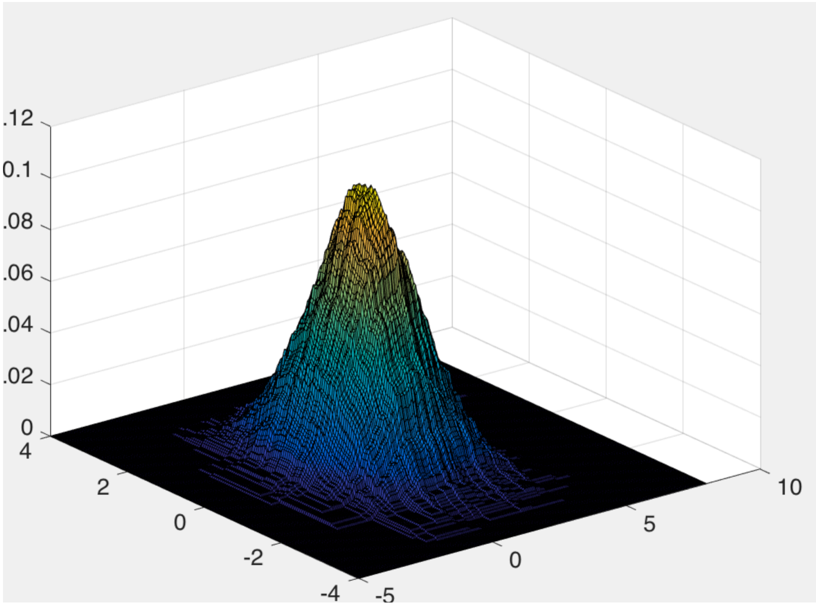
```
>> err_rate

err_rate =

    0.1988
```

2. 用高斯 parzen 窗和矩形窗分别估计已知 label 的一个 component 的条件分布，这是其中第二个 component 估计的结果：





这里用矩形 parzen 窗估计两个条件分布，数值计算得到的分类错误率为：0.2063。

用高斯 parzen 窗估计两个条件分布，数值计算得到的分类错误率为：0.2049。而且从图中可以看出样本数量不足够多时，高斯估计能够得到更平滑的概率分布。

3. 实现了一个做 GMM 的 EM 算法 em_gmm 中。

Cal_error 脚本用 p2.m 和 p3.m 中计算的概率 mesh 计算三种算法估计概率的平方误差。EM, Gauss parzen 和 rect parzen 分别估计条件分布得到的均方错误分别为：

	EM	高斯 parzen	矩形 parzen
Component 1	5.8386e-04	0.0011	0.0014
Component 2	4.3785e-04	7.3388e-04	8.1682e-04

很明显可以看到由于 EM 估计用了更多的先验知识（已知分布是混合高斯分布），所以可以在不用调参数(sigma/矩形窗大小)的情况下达到较好的估计分布效果。

5. 使用三种方法分别运行 RUN=20 次，计算用于做分类的 error 的期望和方差如下：

	EM	高斯 parzen	矩形 parzen
error 期望	0.2114	0.2127	0.2140
error 方差	1.2804e-04	4.3271e-05	4.3065e-05

可以看出由于 EM 算法在样本数不多的时候，能够更好的估计概率密度函数，分类错误的期望比较小。但是 EM 算法可能有由于初值没选取好卡在 local maxmimum 的情况，所以其 error 的方差稍大于另外两种基于非参估计方法。

在我的算法里，initialize 的方法如下：initialization 中 sigma_e 为了防止奇异，加入了一个小的正则。

```
% Initialize parameters randomly
pc_e = rand(1, num_comp);
pc_e = pc_e / sum(pc_e); % normalize

u_e = randn(dim, num_comp);

epsilon = 0.1;
sigma_e = zeros(dim, dim, num_comp);
for y=1:num_comp
    Q = rand(dim, dim);
    % avoid singular initialization
    sigma_e(:, :, y) = Q' * Q + epsilon * eye(dim);
end
```

Programming 2

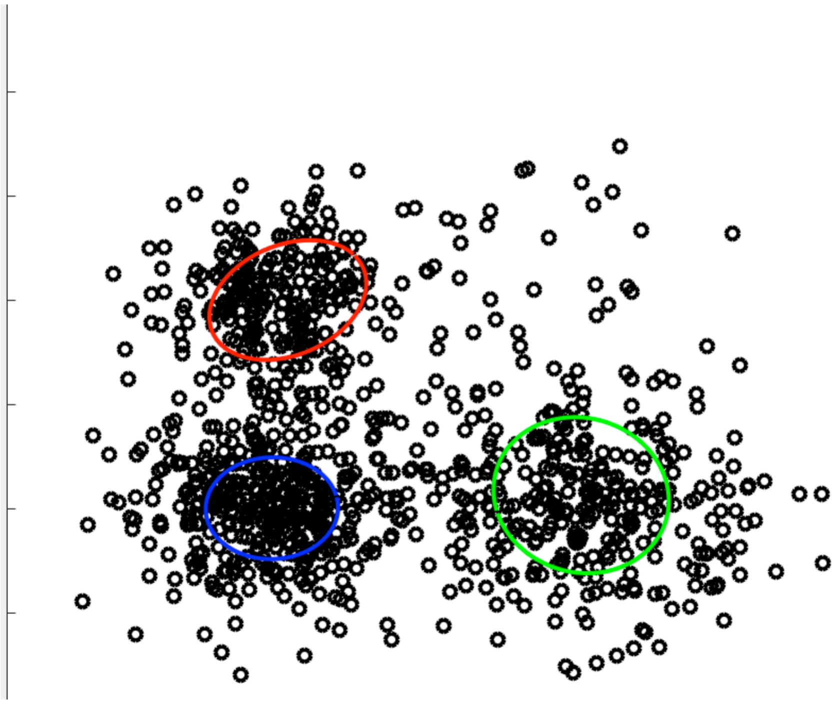
1. 当每个 component 的协方差矩阵分别估计和固定所有 component 的协方差矩阵相同时时，对 m=2…5 几种情况运行 em_mix 程序（每种配置跑了 RUN=5 次），得到结果如下：

Best LL(*10e3)	m=2	m=3	m=4	m=5
协方差矩阵不同	-4.2311	-4.1884	-4.0890	-4.0940
协方差矩阵相同	-4.2583	-4.1724	-4.1592	-4.1267

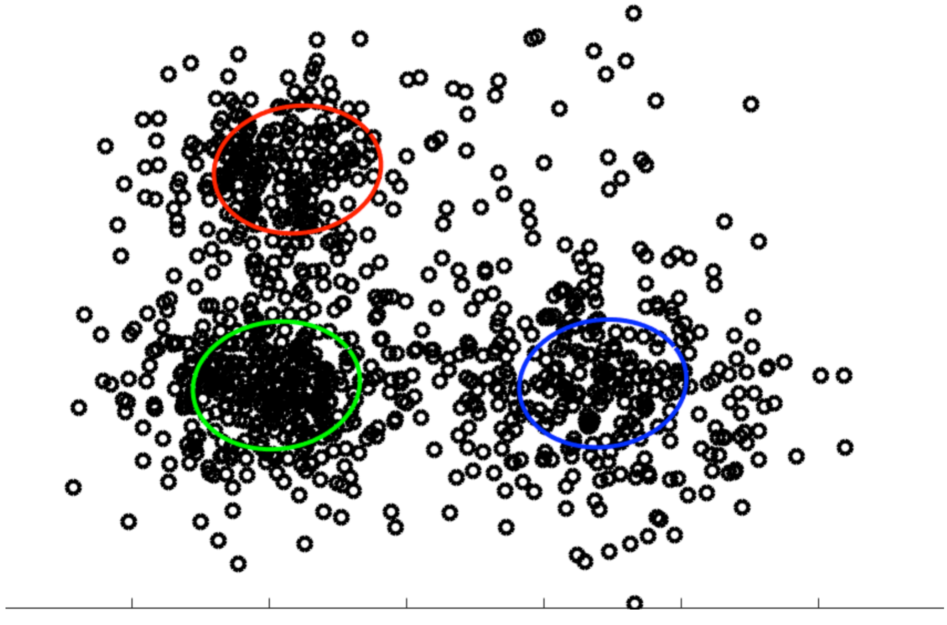
如果使用 BIC 信息准则做 model selection:

BIC(*10e3)	m=2	m=3	m=4	m=5
协方差矩阵不同	8.5181	8.3504	8.2969	8.3346
协方差矩阵相同	8.5572	8.4007	8.3917	8.3395

各个 component 的协方差矩阵不同时的一个拟合结果：



下面是各个 component 的协方差矩阵相同时的 m=3 的一个拟合结果:



使用 best LL 做 model selection 会倾向于选 m 更大的模型，因为只考虑了 likelihood，当模型的 m 越大时，参数越多，表示能力越强，所以可以在训练数据上达到较高的 likelihood。从第一个表格中可以看出两种情况分别是 m=4 和 m=5 的时候 Best LL 最大。当加入 BIC 信息准则(惩罚参数多的模型)做 model selection，会发现倾向于选择参数数量少的模型，但是在这里由于训练数据量比较小。在比较中基本没有起到作用…

从视觉直观来看，确实是三个 components 的时候拟合的最好。