IDS ASSIGNMENT - 3

KANIKI BHIKSHAPATHI
24489

Question - 1

We have a k component Gaussian mixture model

Given that $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \sigma^2 I$

To prove :- EM algorithm coincides with k-means clustering as $\sigma \to 0$

To complete data likelihood function is

$$L(\theta; x, z) = P(x, z/\theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} \left(\tau_j f(x_i, u_j, \Sigma_j) \right)^{I_{z_i = j}}$$

$u_j, \Sigma_j \to$ parameters

$\tau_j = P(z_i = j)$

we have $\sum_{j=1}^{k} \tau_j = 1$

The likelihood function transform to

$$L(\theta; x, z) = \exp\left\{ \sum_{i=1}^{n} \sum_{j=1}^{k} I(z_i = j) \left[\log \tau_j - \frac{1}{2} \log |\Sigma_j| - \frac{1}{2} (x_i - u_j)^T \Sigma_j^{-1} (x_i - u_j) - \frac{d}{2} \log (2\pi) \right] \right.$$

Let

$$T_{j,i} = P\left[z_i = j \mid x_i = x_i ; \theta^{(t)} \right]$$

$$= \frac{P\left[x_i \mid z_i = j ; \theta^{(t)} \right] . P\left[x_i, z_i ; \theta^{(t)} \right]}{\sum_j P\left[x_i \mid z_i = j ; \theta^{t} \right] P\left[x_i ; z_i ; \theta^{t} \right]}$$

using Bayes theorem

$$T_{j,i} = \frac{\tau_j . f(x_i, z, \theta^{(t)})}{\sum_{j=1}^{k} \tau_j f(x_i, z_i ; \theta^{(t)})}$$

$$f(x_i, z_i ; \theta^{(t)}) = \exp\left[-\frac{1}{2} (x_i - u_j)^T \Sigma_j^{-1} (x_i - u_j) \right] \times \frac{1}{(2\pi |\Sigma_j|)^{d/2}}$$

$$T_{j,i} = \frac{\tau_j \times \exp\left[-\frac{1}{2} (x_i - u_j)^T \Sigma_j^{-1} (x_i - u_j) \right] \times \frac{1}{(2\pi |\Sigma_j|)^{d/2}}}{\sum_{j=1}^{k} \tau_j \times \exp\left[-\frac{1}{2} (x_i - u_j)^T \Sigma_j^{-1} (x_i - u_j) \right] \times \frac{1}{(2\pi |\Sigma_j|)^{d/2}}}$$

$$= \frac{\pi_j \exp\left[-\frac{1}{2}(x_i-u_j)^T \Sigma_j^{-1}(x_i-u_j)\right]}{\sum_{j=1}^{K} \pi_j \exp\left[-\frac{1}{2}(x_i-u_j)^T \Sigma_j^{-1}(x_i-u_j)\right]}$$

For $j=1$, $i=1$

$$= \frac{T_1}{\sum_{j=1}^{K} T_j \exp\left[\frac{1}{2}(x_1-u_1)^T \Sigma_1^{-1}(x_1-u_1) - \frac{1}{2}(x_1-u_j)^T \Sigma_j^{-1}(x_i-u_j)\right]}$$

$$= \frac{T_1}{T_1 \exp[0] + \sum_{j=2}^{K} T_j \exp\left[\frac{1}{2}(x_1-u_1)^T \Sigma_1^{-1}(x_1-u_1) - \frac{1}{2}(x_1-u_j)^T \Sigma_j^{-1}(x_i-u_j)\right]}$$

$$= \frac{T_1}{T_1 + \sum_{j=2}^{K} T_j \exp\left[\frac{1}{2}\sum_{k=1}^{d}\frac{(x_{1k}-u_{1k})^2}{\sigma^2} - \frac{1}{2}\sum_{k=1}^{d}\frac{(x_{1k}-u_{jk})^2}{\sigma^2}\right]}$$

From previous eqn if point 1 is closer to u_1

i.e $\sum_{k=1}^{d}(x_{1k}-u_{1,k})^2 < \sum_{k=1}^{d}(x_{1k}, u_{jk})^2 \quad \forall j \geq 2$

then as $\sigma \to 0$ point $T_{1,1} = 1$

Otherwise if $\sum_{k=1}^{d}(x_{1,k}-u_{1,k})^2 > \sum_{k=1}^{d}(x_{1k}-u_{jk})^2$

for atleast $1, j \geq 2$

Then $T_{11} \to 0$ as $\sigma \to 0$

Based on $T_{ji} \to \begin{cases} 1 & \text{if } k = \arg\min_j \|x_i - u_j\|^2 \\ 0 & \text{otherwise} \end{cases}$

$$u_j^{t+1} = \frac{\sum_{i=1}^{n} T_{j,i}^{(t)} x_i}{\sum_{i=1}^{n} T_{j,i}^{(t)}} = \frac{\sum_{x_i \in c} x_i}{d_j}$$

d_j ... no of points belonging to that cluster

$$T_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} T_{j,i}^{(t)} = \frac{n_j}{n} \rightarrow \text{no of points assigned to that cluster}$$

The above is same a k-means clustering

Hence as $\epsilon \rightarrow 0$ EM algo to estimates coincides with k-means clustering.

Question - 2

(a) $P(A) = \frac{1}{2}$, $P(B) = u$, $P(C) = 2u$ $P(D) = \frac{1}{2} - 3u$

No of students with A grade = a

No of students with B grade = b

No of students with C grade = c

No of students with D grade = d

$$P(x; u) = \left(\frac{1}{2}\right)^a \cdot (u)^b \cdot (2u)^c \left(\frac{1}{2} - 3u\right)^d$$

$$\text{MLE } \hat{u} = \arg\max_{u} P(x; \theta)$$

$$= \left(\frac{1}{2}\right)^a (u)^b (2u)^c \left(\frac{1}{2} - 3u\right)^d$$

$$\frac{d\hat{u}}{du} = 2^{(c-a)}(b+c)\, u^{b+c-1} \left(\frac{1}{2} - 3u\right)^d + (-3d)\left(\frac{1}{2} - 3u\right)^{d-1} 2^{c-a} \cdot u^{b+c} = 0$$

$$\left(\frac{b+c}{2}\right)\left(\frac{1}{2} - 3u\right) - 3du = 0$$

$$u = \frac{b+c}{6(b+c+d)}$$

$$\boxed{\hat{u} = \frac{b+c}{6(b+c+d)}}$$

(b) Given $\hat{u}$ the current estimate of u we have

$P(A) = \frac{1}{2}$, $P(B) = \hat{u}$, $P(C) = 2\hat{u}$, $P(D) = \frac{1}{2} - 3\hat{u}$

Let $\hat{a}$ & $\hat{b}$ be the expected no of students that has graded A & B respectively.

h is the number of student that have grade A or B

A student can have either grade A or grade B.

$P(A \cup B) = P(A) + P(B) = \frac{1}{2} + \hat{u}$ (A & B are mutually exclusive event)

Total no of students who belong to $A \cup B$ are h.

$\hat{a} = h \cdot P(A/_{A \cup B})$ $\quad \hat{b} = h \cdot P\left(\frac{B}{A \cup B}\right)$

$\hat{a} = h \times \frac{0.5}{0.5 + \hat{u}}$ $\quad \hat{b} = \frac{h \times \hat{u}}{0.5 + \hat{u}}$

Let z_i denote the group of ith student.

i.e z_i can take value $\{A, B, C, D\}$ Let $A=1, B=2, C=3, D=4$

$\therefore$ Complete likelihood function will be

$$L(x, z; \theta) = \prod_{i=1}^{h+c+d} \prod_{j=1}^{4} \left(T_j f(x_i | z_i)\right)^{I(z_i = j)}$$

$$= \left(\frac{1}{2}\right)^{\hat{a}} \cdot (u)^{\hat{b}} \cdot (2u)^{c} \left(\frac{1}{2} - 3u\right)^{d}$$

$T_j = P[z_i = j]$

$P(x_i \in C/_{z_i=3}) = 1$

$P(x_i \in D | z_i = 4) = 1$

$P(x_i \in B | z_i = 2) = 1$

$P(x_i \in A | x_i = 1) = 1$

Log Likelihood

$$\ln(x, z; \theta) = \hat{a} \ln\left(\frac{1}{2}\right) + \hat{b} \ln u + C \ln(2u) + d \ln\left(\frac{1}{2} - 3u\right)$$

To maximize log likelihood we take derivative

$$\therefore \frac{\partial \ln L}{du} = 0 + \frac{\hat{b}}{u} + \frac{c}{u} + \frac{d(-3)}{\frac{1}{2} - 3u} = 0$$

$$\frac{\hat{b} + c}{u} = \frac{3d}{\frac{1}{2} - 3u} \Rightarrow u = \frac{\hat{b} + c}{6(\hat{b} + c + d)}$$

Here $\hat{b} = \frac{h\hat{u}}{0.5 + \hat{u}}$, $\hat{a} = \frac{h \times 0.5}{0.5 + \hat{u}}$

$u^{t+1} = \frac{\hat{b} + c}{6(\hat{b} + c + d)}$ $\quad \hat{b} \rightarrow$ expected no of students who graded B out total

Question - 3

Let $y = \beta_0 + \beta_1 x$ be the best fit line

To find best fit line we minimize Least square error

$$\therefore \min \sum_{i=1}^{30} (y_i - \beta_0 - \beta_1 x_i)^2 = L$$

Differentiate w.r.t to β_0 & set it to 0

$$\frac{\partial L}{\partial \beta_0} = \sum_{i=0}^{30} 2(y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

$$\sum_{i=1}^{30} y_i - \beta_0 \sum_{i=0}^{30} 1 - \beta_1 \sum_{i=0}^{30} x_i = 0$$

$$\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

$$\frac{dL}{d\beta_1} = \sum_{i=0}^{30} 2(y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

$$= \sum_{i=0}^{30} y_i x_i - \beta_0 \sum_{i=1}^{30} x_i - \beta_1 \sum_{i=1}^{30} x_i^2 = 0$$

$$= \sum_{i=0}^{30} y_i x_i - (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{30} x_i - \beta_1 \sum_{i=1}^{30} x_i^2 = 0$$

$$= \sum_{i=0}^{30} y_i x_i - 30 \bar{x} \bar{y} = \beta_1 \sum_{i=1}^{30} x_i^2 - \beta_1 30 \bar{x}^2$$

$$\beta_1 = \frac{\left[\sum_{i=1}^{30} \frac{y_i x_i}{30} - \bar{x}\bar{y}\right]}{\sum_{x=1}^{30} \frac{x_i^2}{30} - \bar{x}^2} = \beta_1 = \frac{\rho_{xy} \times \sigma_y \times \sigma_x}{\sigma_x^2}$$

$$\boxed{\beta_1 = \rho_{xy} \times \frac{\sigma_y}{\sigma_x}}$$

$$\beta_0 = \bar{y} - \left(\rho_{xy} \cdot \frac{\sigma_y}{\sigma_x}\right) \bar{x}$$

$$y = \bar{y} - \left(\rho_{xy} - \frac{\sigma_y}{\sigma_x}\right) \bar{x} + \rho_{xy} \times \frac{\sigma_y}{\sigma_x} \cdot x$$

$$y = 2.0667 + 0.2333x$$

The points $(\bar{x}, \bar{y})$ does satisfy the above eqn.

Question - 4

Given $\rightarrow$ $y_i \sim N(\beta_0 + x_i^T\beta, \sigma^2)$ $i = 1, 2, \ldots, N$ & $\beta \in R^p$

$\beta_j \sim N(0, \tau^2)$ $\forall j = 1, \ldots, p$

log posterior density of β $p(\beta|x)$

$u = \beta_0 + x_i^T\beta$ $E(u) = \beta_0$

$$P(\beta) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\tau}} \exp\left(-\frac{\beta_j^2}{2\tau^2}\right)$$

using Bayes theorem

$$P(\beta|(x_i, y_i)) = \frac{P(x, y|\beta)P(\beta)}{\int P(x, y|\beta)P(\beta)\,d\beta}$$

$$\ln p(\beta|(x,y)) \propto \ln(p((x,y))|\beta) + \ln(P|(\beta))$$

$$\propto \ln\left(\prod_{i=1}^{N} \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \beta_0 - x_i^T\beta}{\sigma^2}\right)^2\right]\right)$$

$$+ \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\tau}} \exp\left[\frac{-\beta_j^2}{2\tau^2}\right]\right)$$

$$\ln p(\beta|(x,y)) \propto \ln c - \sum_{i=1}^{N} \frac{1}{2}\left(\frac{y_i - \beta_0 - x_i^T\beta}{\sigma^2}\right)^2 + \sum_{i=1}^{N}\left(\frac{-\beta_j^2}{2\tau^2}\right)$$

$$\boxed{-\ln P(\beta|(x,y)) \times 2\sigma^2 \propto \sum_{i=1}^{N}(y_i - \beta_0 - x_i^T\beta)^2 + \frac{\sigma^2}{\tau^2}\sum_{i=1}^{N}\beta_i^2}$$

Question - 5

(a) Sol^n :- For loss function L to be appropriate

① The loss $L(yF(x))$ must decrease as $yF(x)$ increases because +ve margins $yF(x)$ indicate more confident & correct predictions

② It must always be non-negative i.e $L(yF(x)) \geq 0$

Based on the above we can say that (c), (d), & (e) are not the right appropriate loss function.

(b) The loss for plot (b) is capped at 1 because of which the effect of outliers is limited. The plot (a) function is very sensitive to outliers at it keep on increasing as $yF(x)$ decreases.

(c) Gradient of L w.r.t $F(x)$

Let $z = yF(x)$

$$L(z) = \frac{1}{1+\exp(z)} \qquad = \frac{-1+\exp(z)}{(1+\exp(z))^2}$$

$$F(u) = w_0 + \sum_{j=1}^{d} w_j x_j$$

$$\frac{dL}{dw_0} = \frac{dL}{dF(x)} \cdot \frac{dF(x)}{d(w_0)} \quad \text{—(1)}$$

$$\frac{dL}{dw_0} = -\frac{\exp(yF(x)).y}{(1+\exp(yF(x))^2}$$

Similarly $w_j \Rightarrow \frac{\partial L}{dw_j} = -\frac{\exp(yF(x))\, y.x_j}{[1+\exp(yF(x))]^2}$

Hence the update rules will be

$$w_0^{k+1} = w_0^k + \eta . \frac{\exp(yF(x))\, y}{(1+\exp(yF(x))^2}$$

$$w_j^{k+1} = w_j^k + \frac{\eta \exp(yF(x))\, y.x_j}{(1+\exp(yF(x)))^2}$$