

24/03/2020

Mixture model

$$x \sim f(x|\theta) \quad f(x|\theta) = \sum_{k=1}^K \pi_k \varphi(x; \mu_k, 1)$$

$$\varphi(x; \mu_k, 1) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(x - \mu_k)^2\right\}$$

$$\theta = (\mu_1, \mu_2, \dots, \mu_K, \pi_1, \dots, \pi_{K-1})$$

$$\sum_k \pi_k = 1$$

$$L(x_1, \dots, x_n | \theta) = \prod_{i=1}^n \sum_{k=1}^K \pi_k \varphi(x_i; \mu_k, 1) \quad \text{likelihood}$$

$$\ell(x_1, \dots, x_n | \theta) = \sum_{i=1}^n \log\left(\sum_{k=1}^K \pi_k \varphi(x_i; \mu_k, 1)\right) \quad \text{loglik}$$

$$\frac{\partial \ell}{\partial \mu_e} = \sum_{i=1}^n \frac{1}{\sum \pi_k \varphi(x_i; \mu_k, 1)} \cdot \pi_e \cdot \varphi(x_i; \mu_e, 1) \cdot 2(x_i - \mu_e) = 0$$

- $\gamma_e(e)$

Hard to solve

(*)

Our data $x_1 \dots x_n$

Augmented data $z_1 \dots z_n$

$$z_i = k \Leftrightarrow x_i \sim \varphi(x; \mu_{k,1})$$

Augmented data set $(\bar{x}, \bar{z})$

If $\bar{z}$ was known

$$\begin{aligned} f(x, z | \theta) &= f(x | z, \theta) \cdot f(z | \theta) = \\ &= \underbrace{\sum_{k=1}^2 \pi_k f(x | \mu_{k,1}) \cdot \underline{\underline{1(z=k)}}}_{\text{}} \end{aligned}$$

$$L(x, z | \theta) = \prod_{i=1}^n \sum_{k=1}^2 \pi_k \varphi(x_i; \mu_{k,1}) \cdot 1(z_i = k) =$$

$$= \underbrace{\prod_{i: z_i=1} \pi_1 \varphi(x_i; \mu_{1,1})}_1 \times \underbrace{\prod_{i: z_i=2} \pi_2 \varphi(x_i; \mu_{2,1})}_2$$

$$\begin{aligned} \ell(x, z | \theta) &= \sum_{i: z_i=1} (\log \pi_1 + \log \varphi(x_i; \mu_{1,1})) + \\ &\quad \sum_{i: z_i=2} (\log \pi_2 + \log \varphi(x_i; \mu_{2,1})) = \end{aligned}$$

$$= \sum_{i=1}^n \sum_{k=1}^2 \left[\log \pi_k + \log \varphi(x_i; \mu_{k,1}) \right] \times \mathbb{1}(Z_i = k)$$

If Z was known we could solve this problem!



We do not know Z but what is his posterior distr.?

$$\underbrace{f(Z_i = e | x_i)}_{P(Z_i = e | x_i)} = \frac{\underbrace{f(x_i | Z_i = e)}_{f(x_i)} \cdot \underbrace{f(Z_i = e)}}{f(x_i)} =$$

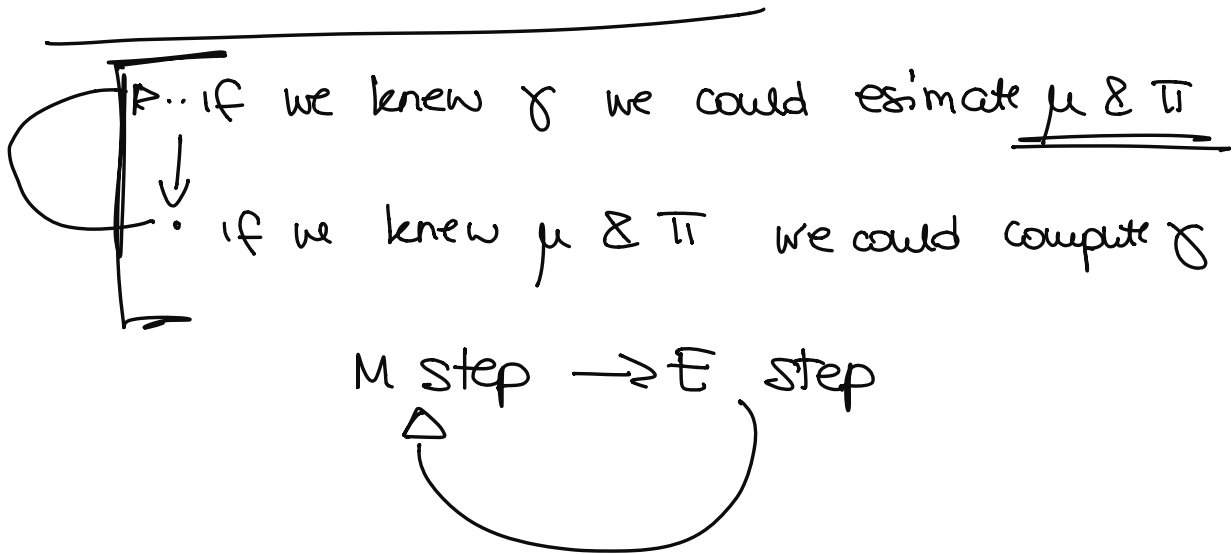
$$= \frac{\varphi(x_i; \mu_{e,1}) \pi_e}{\sum_k \pi_k \varphi(x_i; \mu_{k,1})} = \gamma_{Z_i}(e)$$

Look back (*)

$$\frac{\partial \ell}{\partial \mu_e} = \sum_{i=1}^n \gamma_{Z_i}(e) \cdot (x_i - \mu_e) = 0$$

. If we knew $\gamma_{Z_i}(e)$ we could estimate μ_e
(and also π_e)

$$\hat{\mu}_e = \frac{\sum \gamma_{zi}(e) \cdot x_i}{\sum \gamma_{zi}(e)}$$



What is $Q(\theta)$ for the gaussian mixture problem?

$$\theta = (\mu_1, \mu_2, \pi_1)$$

$$\begin{aligned}
 E(\ell(x, z | \theta^{(j)}, x)) &= \\
 &= E \left[\sum_{i=1}^n \sum_{k=1}^2 \left\{ \log \pi_k^{(j)} + \log \varphi(x_{ij} | \mu_k^{(j)}, 1) \right\} \cdot \underline{1(z_i=k)} \mid \theta^{(j)}, x \right] \\
 &= \sum \sum \left[\log \pi_k^{(j)} + \log \varphi(x_{ij} | \mu_k^{(j)}, 1) \right] E \left[\underline{z_i=k} \mid \theta^{(j)}, x \right]
 \end{aligned}$$

$$E[z_i=k | \theta^{(j)}, x] = P(z_i=k | \theta^{(j)}, x) =$$

$$Z_i \in \{1, 2\}$$

$$\pi(Z_i=1) = \pi_1 \quad \pi(Z_i=2) = \pi_2$$

$$= \frac{\prod \frac{\phi(x_i | Z_i=k, \theta^{(i)}) \cdot P(Z_i=k | \theta^{(i)})}{\sum \phi(x_i | \theta^{(i)})}} = \delta_{Z_i}(k)$$

As we found earlier

Gene examples



$$p_T = P(\text{having gene } T)$$

$$p_t = P(\text{---} \quad \text{---} \quad t)$$

- Want to estimate p_T, p_t $\theta = (p_T, p_t)$

- Data $n_D \rightarrow$ # with dark eyes
 $n_L \rightarrow$ --- light eyes

- observed data set $X = (n_D, \underline{n_L})$

- expended data set $Y = (\underline{n_{TT}}, \underline{n_{Tt}}, \underline{n_{tt}})$

$$n_{TT} + n_{Tt} = n_D$$

$$n_{tt} = n_L \quad n = n_D + n_L$$

$$Y|\theta \sim \text{multinomial}(n, (p_T^2, 2p_T p_t, p_t^2))$$

$$\log L(\theta|Y) = n_{TT} \log p_T^2 + n_{Tt} \log 2p_T p_t + n_{tt} \log p_t^2 + \underbrace{\log \binom{n}{n_{TT} n_{Tt} n_{tt}}}_{\text{const wrt } \theta}$$

E step Compute $E[\log L(Y|\theta) | x, \theta^{(t)}] = Q$

M step maximise Q wrt θ

• Note that when n_D (n_{TT}, n_{Tt}) is ~~multivariate~~ ^{nominal}

distributed with para $\left(n_D, \left(\frac{p_T^2}{2p_T p_t + p_T^2}, \frac{2p_T p_t}{2p_T p_t + p_T^2} \right) \right)$

E step

$$E[\log L(\theta | y) | x, \theta^{(i)}]$$

$$E(n_{TT} | x, \theta^{(i)}) = n \cdot p = n_D \cdot \underbrace{\frac{p_T^2}{2p_T p_t + p_T^2}}_{n_{TT}^{(i)}}$$

$$E(n_{tT} | x, \theta^{(i)}) = n_D \frac{2p_T^{(i)} p_t^{(i)}}{\underbrace{2p_T^{(i)} p_t^{(i)} + p_T^{(i)2}}_{n_{TT}^{(i)}}}$$

$$E(n_{tt} | x, \theta^{(i)}) = n_{tt}$$

$$x = (n_D, \underbrace{n_{TT}}_{n_{tt}})$$

so we have that

$$Q(\theta | \theta^{(i)}) = n_{TT}^{(i)} \cdot \log p_T^2 + n_{tT}^{(i)} \cdot \log 2 p_t p_T + n_{tt} \log p_t^2 + \text{constant wrt } \theta =$$

$1 = p_T + p_t$

$$= 2 n_{TT}^{(i)} \cdot \log p_T + n_{tT}^{(i)} \log 2 p_T + (n_{tT}^{(i)} + 2 n_{tt}) \log (1 - p_T)$$

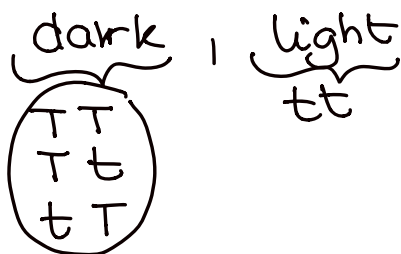
M STEP $\rightarrow$ $\odot$ Maximise $Q(\theta | \theta^{(i)})$ wrt θ

$$\theta = p_T$$

$$\frac{\partial Q}{\partial p_T} = \frac{2n\pi^{(i)}}{p_T} + \frac{n\tau^{(i)}}{2p_T} \cdot 2 + \frac{n\tau^{(i)} + 2n_{tt}(-1)}{1-p_T} = 0$$

~~2104~~ $\hat{p}_T = \frac{2n\pi^{(i)} + n\tau^{(i)}}{2n}$

What is the incomplete likelihood?



dark $\sim$ binomial(n , $p_T^2 + 2p_T p_t$)

$$\log L(x | \theta) \propto n_D \cdot \log(p_T^2 + 2p_T p_t) + (n - n_D) \log(1 - p_T^2 - 2p_T p_t)$$