CS771 Assignment Abhay

Solution 1

Suppose $\vec{v}$ is an eigen vector of matrix $S' = \frac{1}{N} XX^T$. By definition, $S'v = \lambda v$, where λ is corresponding eigen value

$$\therefore \quad \frac{1}{N}(XX^T)v = \lambda v$$

Premultiplying by X^T both sides, we get

$$\frac{1}{N} X^T X X^T v = \lambda (X^T v).$$

substituting $X^T v = u$ $[X^T \to D \times N, v \in \mathbb{R}^N \therefore u \in \mathbb{R}^D]$.

$$\frac{1}{N}(X^T X) u = \lambda (u)$$

From above expression, it is clear that $\underline{u = X^T v}$ is eigen vector for $\frac{1}{N}(X^T X)$ ie S.

In normal way, time complexity to compute eigen vector of S is $O(KD^2) \longrightarrow T_1$

For this way, time complexity will be

$$\underbrace{O(KN^2)}_{\text{eigen vectors for } S'} + \underbrace{O(KND)}_{\text{Matrix Multiplication}} \longrightarrow T_2$$

Since, we are given $D > N$ in the question, thus $T_1 > T_2$.

i.e we can say that computing eigen vector of S through S' is more efficient than computing them directly.

Solution 2.

Given: Poisson distribution, N webserves monitored for M minutes.

$k_{n,m} \rightarrow$ no. of hits to n^{th} webserver in minute m.

As shown in the hint, the complete data likelihood is as follows.

$$p(k, z | \lambda, \pi) = \prod_{n=1}^{N} \prod_{l=1}^{L} \left[p(z_n = l) \prod_{m=1}^{M} \text{Poisson}(k_{n,m} | \lambda_l) \right]^{1[z_n = l]}.$$

$$1[z_n = l] = \begin{cases} 1 & \text{if } z_n = l \\ 0 & \text{otherwise} \end{cases}.$$ and also $p(z_n = l) = \pi_l$.

Taking log both sides.

$$CLL = \sum_{n=1}^{N} \sum_{l=1}^{L} 1[z_n = l] \left[\log(\pi_l) + \log\left(\prod_{m=1}^{M} \text{Poisson}(\ldots) \right) \right].$$

$$= \sum_{n=1}^{N} \sum_{l=1}^{L} z_{nl} \left[\log(\pi_l) + \log\left(\prod_{m=1}^{M} \frac{1}{e^{\lambda_l}} \frac{\lambda_l^{k_{n,m}}}{(k_{m,m})!} \right) \right.$$

$$= \sum_{n=1}^{N} \sum_{l=1}^{L} z_{nl} \left[\log(\pi_l) + \sum_{m=1}^{M} \left(k_{n,m} \log \lambda_l - \lambda_l - \log((k_{n,m})!) \right) \right]$$

Thus, the complete data log likelihood is given as.

$$\sum_{n=1}^{N} \sum_{l=1}^{L} z_{nl} \left[\log(\pi_l) + \sum_{m=1}^{M} \left(k_{n,m} \log \lambda_l - \lambda_l - \log(k_{n,m}!) \right) \right]$$

$z_{nl} = 1$ and all other components of one hot vector $z_n = 0$.

Estimating z_{nl}. $\rightarrow$ E-step

$$E[z_{nl}] = 1 \times p(z_{nl}=1|k_n,\theta) + \cancel{0 \times p(z_{nl}=0|k_n)}$$

$$= p(z_{nl}=1|k_n)$$

$$\propto p(z_{nl}=1)\, p(k_n|z_{nl}=1,\theta)$$

$$\propto \cancel{\pi_k \mathcal{N}(\ldots)} \cdot \pi_l \left[\prod_{m=1}^{M} \text{Poisson}(k_{nm}|\lambda_l)\right]$$

$$\propto \pi_l \left(\frac{e^{-\lambda_l M} \lambda^{\sum_{m=1}^{M} k_{nm}}}{\prod_{m=1}^{M} (k_{nm}!)} \right)$$

c.) M-step.

Taking $E[z_{nl}] = \gamma_{nl}$.

Expected complete Data log-likelihood is.

$$\mathcal{EL} = \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl} \left[\log(\pi_l) + \sum_{m=1}^{M} \left(\log(\lambda_l) . k_{n,m} - \lambda_l - \log(k_{n,m}!) \right) \right]$$

Taking derivative w.r.t λ_l., ($\gamma_{ni} = 0 \ \forall i \neq l$)

$$\frac{\partial \mathcal{EL}}{\partial \lambda_l} = \sum_{n=1}^{N} \gamma_{nl} \left[\cancel{\frac{\partial}{\partial \lambda_l} \log(\pi_l)} + \frac{\partial}{\partial \lambda_l} \left(\sum_{m=1}^{M} \left(k_{n,m} \log \lambda_l - \lambda_l - \log(\ldots) \right) \right) \right]$$

$$= \sum_{n=1}^{N} \gamma_{nl} \left[\sum_{m=1}^{M} \left(\frac{k_{n,m}}{\lambda_l} - 1 \right) \right]$$

Now, equating it with 0., we get

$$\sum_{n=1}^{N} \gamma_{nl} \left[\sum_{m=1}^{M} \left(\frac{k_{n,m}}{\lambda_l} - 1 \right) \right] = 0$$

$$\Rightarrow \sum_{n=1}^{N} \left(\gamma_{nl} \sum_{m=1}^{M} \frac{k_{n,m}}{\lambda_l} - \gamma_{nl} M \right) = 0$$

$$\Rightarrow \sum_{n=1}^{N} \gamma_{nl} \sum_{m=1}^{M} \frac{k_{n,m}}{\lambda_l} = \sum_{n=1}^{N} \gamma_{nl} M$$

$$\Rightarrow \boxed{\lambda_l = \frac{\sum_{n=1}^{N} \gamma_{nl} \sum_{m=1}^{M} k_{n,m}}{\sum_{n=1}^{N} \gamma_{nl} M}}$$ $\rightarrow$ where $\gamma_{nl} = E[z_{nl}]$ ✓

Estimating π_l

indep. of π_l (in previous page)

E[CLL] $\swarrow$ $E.L = \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl} \left[\log \pi_l + \sum_{m=1}^{M} (\text{some terms}) \right]$

We need to max. above eqn, but there is constraint $\sum_{l=1}^{L} \pi_l = 1$, so, we will use Langrangean method,

$$\ell = \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl} \log(\pi_l) + \alpha * \left(1 - \sum_{l=1}^{L} \pi_l\right).$$ [ignored ind. terms]

$$\frac{\partial \ell}{\partial \pi_l} = \sum_{n=1}^{N} \frac{\gamma_{nl}}{\pi_l} - \alpha$$, Equating to zero gives us.

$$\sum_{n=1}^{N} \frac{\gamma_{nl}}{\pi_l} = \alpha \Rightarrow \pi_l = \frac{\sum_{n=1}^{N} \gamma_{nl}}{\alpha}.$$, but $\alpha = ?$

$$\frac{\partial \ell}{\partial \alpha} = 0 \Rightarrow \sum_{l=1}^{L} \pi_l = 1 \Rightarrow \frac{\sum_{l=1}^{L} \sum_{n=1}^{N} \gamma_{nl}}{\alpha} = 1$$

Hence $\alpha = \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl}$. , therefore

$$\boxed{\pi_l = \frac{\sum_{n=1}^{N} \gamma_{nl}}{\sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl}}}$$ ✓

Final Ans

$\Downarrow$

Now since $\sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl} = \underline{N}$. $\therefore$ $\pi_l = \frac{\sum_{n=1}^{N} \gamma_{nl}}{N}$

Solution 3. $\boxed{\text{Part 1}}$

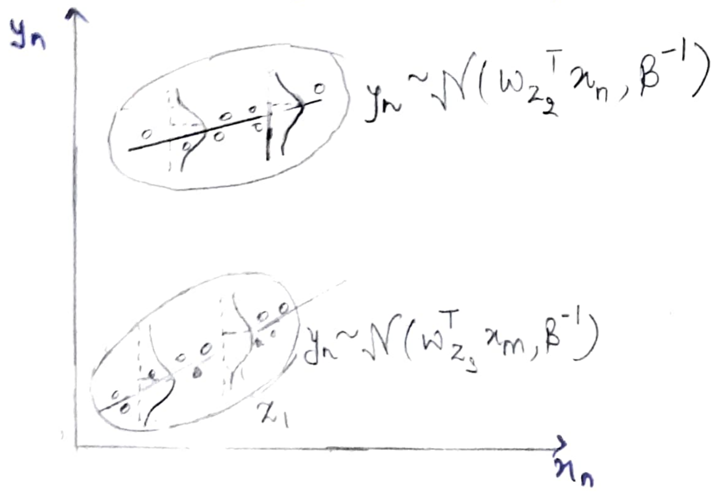
Generative story as given in question.

(i) $z_n \sim$ multinomial $(\pi_1, \pi_2, \ldots \pi_K)$

(ii) Generate inputs $x_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$

(iii) Outputs $y_n \sim \mathcal{N}(w_{z_n}^T x_n, \beta^{-1})$.

→



Latent variable model

The model will learn a combination of k-linear regressions, as depicted by the above graph. In resp. cluster, the input-output relationship is linear. However, in standard linear regression, it will only learn a single linear boundary. Also, the LVM model will separate the outliers in clustering, therefore also help in reducing their effect on predictions of model.

Now A.T.Q. $p(z_n|\theta) = \text{multinoulli}(\pi_1, \ldots \pi_K)$

$$p(x_n|z_n, \theta) = \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$$

$$p(y_n|x_n, z_n, \theta) = \mathcal{N}(w_{z_n}^T x_n, \beta^{-1}).$$

3.1.2 Deriving EM algorithm.

$$\cancel{CLL = \prod} \qquad CLL = \sum_{n=1}^{N} \log p(x_n, y_n, z_n | \theta)$$

$$= \sum_{n=1}^{N} \log \left[p(y_n | x_n, z_n, \theta)\, p(x_n | z_n, \theta)\, p(z_n | \theta) \right]$$

$$= \sum_{n=1}^{N} \log p(y_n | x_n, z_n, \theta) + \log p(x_n | z_n, \theta) + \log (z_n | \theta).$$

$$= \cancel{\sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log(\mathcal{N}(\mu_k, \Sigma_k)) + \log \pi_k + \log \mathcal{N}(y_n | w \right.}$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \mathcal{N}(w_k^T x_n, \beta^{-1}) + \log \mathcal{N}(\mu_k, \Sigma_k) + \log (\pi_k) \right]$$

E-step.

$$p(z | x, Y, \theta) = \prod_{n=1}^{N} p(z_n | x_n, y_n, \theta)$$

$$p(z_n | x_n, y_n, \theta) = \frac{p(y_n | x_n, z_n, \theta)\, p(x_n | z_n, \theta)\, p(z_n | \theta)}{p(x_n, y_n | \theta)} \rightarrow \text{Independent of } z.$$

$$p(z_n = k | x_n, y_n, \theta) \propto \mathcal{N}(w_k^T x_n, \beta^{-1})\, \mathcal{N}(\mu_k, \Sigma_k)\, \pi_k.$$

$$p(z_n = k | x_n, y_n, \theta) = \frac{\mathcal{N}(w_k^T x_n, \beta^{-1})\, \mathcal{N}(\mu_k, \Sigma_k)\, \pi_k}{\sum_{l=1}^{K} \mathcal{N}(w_l^T x_n, \beta^{-1})\, \mathcal{N}(\mu_l, \Sigma_l)\, \pi_l}.$$

$$\gamma_{nk} = E[z_{nk}] = 1 \times p(z_n = k) = \text{same as above (")}.$$

M-step:

$$E[CLL] = \sum_{n=1}^{N} \sum_{k=1}^{K} E[z_{nk}] \left(\log \pi_k + \log \mathcal{N}(\mu_k, \Sigma_k) + \log \mathcal{N}(w_k^T x_n, \beta^{-1}) \right)$$

For maximization, we will differentiate this w.r.t to. μ_k, Σ_k and. π_k and w_k.

This will be same as done in class. for GMM.

$$\hat{\pi}_k = \frac{N_k}{N} \qquad \hat{\Sigma}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} (x_n - \hat{\mu}_k)(x_n - \hat{\mu}_k)^T}{N_k}$$

$$\hat{\mu}_k = \frac{\sum_{n=1}^{N} \gamma_{nk} x_n}{N_k} \qquad \underbrace{\hat{w}_k = \left(\sum_n \gamma_{nk} x_n x_n^T \right)^{-1} \left(\sum_n \gamma_{nk} x_n y_n \right)}_{\text{shown on next page}}$$

here $\gamma_{nk} = E[z_{nk}]$, calculated in E step

$N_k = \sum_{n=1}^{N} \gamma_{nk}$ = effective no. of pts in cluster k

---

Overall EM algorithm

1. Initialize $\theta = \theta^0$, set $t=1$
2. E-step :- Compute $E[z_{nk}]$
3. M-step :- Maximize $E[CLL]$ and update parameters as described above.
4. Set $t = t+1$, and go to step 2 if not converged.

Calculation of $\hat{w_k}$,

$$\frac{\partial}{\partial w_k}\left(E[CLL]\right) = 0$$

$$E[CLL] = \sum_{n=1}^{N}\sum_{k=1}^{K} E[z_{nk}]\Big(\underbrace{\log \pi_k + \log \mathcal{N}(\mu_k,\Sigma_k)}_{\text{independent of } w_k} + \log \mathcal{N}(w_k^T x_n, \beta^{-1})\Big) \to ③$$

ignoring constants in expression of ③

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk}\left[\text{indep. terms} + -\frac{\beta}{2}\left(y_n - w_k^T x_n\right)^2\right]$$

$$\frac{\partial E[CLL]}{\partial w_k} = \sum_{n=1}^{N} \beta\, \gamma_{nk}\left(y_n - w_k^T x_n\right) x_n$$

equating to zero, we get

$$\sum_{n=1}^{N} \gamma_{nk}\left(y_n - \hat{w}_k^T x_n\right) x_n = 0$$

$$\sum_{n=1}^{N} \gamma_{nk} x_n y_n = \sum_{n=1}^{N} \gamma_{nk} x_n x_n^T \underbrace{\hat{w}_k}_{\text{doesn't depend on } n.}$$

$$\hat{w}_k = \left(\sum_{n=1}^{N} \gamma_{nk} x_n x_n^T\right)^{-1}\left(\sum_{n=1}^{N} \gamma_{nk} x_n y_n\right)$$

Intution of update eq^n of $\hat{w}_k$

in normal regression $w = \left[\sum_{n=1}^{N} (x_n x_n^T)\right]^{-1} (\sum y_n x_n)$.

Our update eq^n is similar, rather it is specific to every cluster, ie w_k. Note our eq^n only considers points belonging to that cluster. This property is governed by γ_{nk} in the expression

ALT-OPT Algorithm.

Instead of $E[z_{nk}]$, we find $\hat{z}_n$

$$\hat{z}_n = \max_{k \in [1,K]} \pi_k \mathcal{N}(w_k^T x_n, \beta^{-1}) \mathcal{N}(\mu_k, \Sigma_k)$$

since $\pi_k = 1/K$

$$\hat{z}_n = \max_{k \in [1,k]} \mathcal{N}(w_k^T x_n, \beta^{-1}) \mathcal{N}(\mu_k, \Sigma_k)$$

Simply replace γ_{nk} with $\hat{z}_{nk}$ in all the update parameters.

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{n=1}^{N} \hat{z}_{nk} x_n$$

$$\hat{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \hat{z}_{nk} (x_n - \hat{\mu}_k)(x_n - \hat{\mu}_k)^T$$

$$\hat{w}_k = \cancel{\frac{1}{N_k}} \left(\sum_{n=1}^{N} \hat{z}_n x_n x_n^T\right)^{-1} \left(\sum_{n=1}^{N} \hat{z}_n y_n x_n\right).$$

Note here $N_k = \sum_{n=1}^{N} \hat{z}_{nk}$

Overall algorithm ALT-OPT

1. Intialize $\theta = \{ \, , \mu_k , \Sigma_k , W_k \}$ as θ_0, set $t = 1$

2. For each n, compute $\hat{z}_n$

$$\hat{z}_n = \underset{k \in [1,K]}{\operatorname{augmax}} \left[\mathcal{N}_k (x_n \mid \mu_k , \Sigma_k) \, \mathcal{N} . (y_n \mid w_k^T x_n , \beta^{-1}) \right]$$

3. Solve MLE problem using updates given in last page. $\forall k$.

4. Set $t = t + 1$, and go to step 2, if not converged.

Part 2

$x_n \rightarrow$ given (not modeled)

$$\pi_k(x_n) \rightarrow \frac{\exp(\eta_k^T x_n)}{\sum_{l=1}^{K} e^{\eta_l^T x_n}}$$

$$\text{CLL} = \sum_{n=1}^{N} \log p(y_n, z_n \mid x_n, \theta)$$

$$= \sum_{n=1}^{N} \left[\log p(y_n \mid z_n, x_n, \theta) + \log p(z_n \mid x_n, \theta)\right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \mathcal{N}(y_n \mid w_k^T x_n, \beta^{-1}) + \log \pi_k(x_n)\right]$$

$$EL = E[CLL] = \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left[-\frac{\beta}{2}(y_n - w_k^T x_n)^2 + \log \pi_k(x_n)\right]$$

$\boxed{\text{E-step}}$

Calculation of $\gamma_{nk} \rightarrow$?

$$p(z_n = k \mid x_n, y_n, \theta) = \frac{p(z_n = k \mid \theta, x_n)\, p(y_n \mid z_n = k, \theta, x_n)}{\sum_{l=1}^{K} p(z_n = l \mid \theta, x_n)\, p(y_n \mid z_n = l)}$$

$$\propto \underbrace{\pi_k(x_n)}_{\propto e^{\eta_k^T x_n}} \cdot \mathcal{N}(y_n \mid w_k^T x_n, \beta^{-1})$$

$$p(z_n = k \mid x_n, y_n, \theta) = \frac{\mathcal{N}(y_n \mid w_k^T x_n, \beta^{-1})\, e^{\eta_k^T x_n}}{\sum_{l=1}^{K} \mathcal{N}(y_n \mid w_l^T x_n, \beta^{-1})\, e^{\eta_l^T x_n}}$$

$$\gamma_{nk} = E[z_n = k] = 1 \times p(z_n = k \mid x_n, y_n, \theta) + 0$$

$$= p(z_n = k \mid x_n, y_n, \theta)$$

$\boxed{\text{M - step}}$

Update for W_k $\longrightarrow$ remains same as part 1.

$$\hat{W_k} = \left(\sum_{n=1}^{N} \gamma_{nk} x_n x_n^T\right)^{-1} \left(\sum_{n=1}^{N} \gamma_{nk} y_n x_n\right) \quad \forall\, k \in [1, K]$$

Now, since η_k is also a parameter in $E[CLL]$

$$\frac{\partial E[CLL]}{\partial \eta_k} = 0.$$

$$\Rightarrow \frac{\partial}{\partial \eta_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \log \pi_k(x_n) = 0$$

$$\Rightarrow \frac{\partial}{\partial \eta_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk} \left(\eta_k^T x_n - \log \sum_{l=1}^{K} e^{\eta_l^T x_n} \right) = 0$$

$$\Rightarrow \frac{\partial}{\partial \eta_k} \sum_{n=1}^{N} \gamma_{nk} \left(\eta_k^T x_n - \underbrace{\log \sum_{l=1}^{K} e^{\eta_l^T x_n}}_{\text{log of sum}} \right) = 0$$

No, we won't be able to find its closed form solution because of the summation term inside the log. As discussed in class, we will need gradient based optimisation to get the point of estimate.