bonus

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 0

Student Name: Sakshi
Roll Number: 180653
Date: December 19, 2020

QUESTION 1

We've to find the eigenvectors of $\mathbf{S}$:

$$\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X}$$

And we're given the eigenvectors of R:

$$\mathbf{R} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$$

Let's look at the singular value decomposition of $\frac{1}{\sqrt{N}}\mathbf{X}$:

$$\frac{1}{\sqrt{N}}\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T = \sum_{n=1}^{N} \sigma_n \mathbf{u}_n \boldsymbol{v_n}^T$$

Here, $\mathbf{U}$&$\mathbf{V}$ are N*N and D*D orthogonal matrices respectively. Σ is N*D diagonal matrix of singular values. Thus, $\mathbf{S}$ and $\mathbf{R}$ can be written as:

$$\mathbf{S} = \frac{1}{N}\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{\Sigma}^T\mathbf{\Sigma}\mathbf{V}^T = \mathbf{V}\mathbf{\Sigma}_S\mathbf{V}^T$$

$$\mathbf{R} = \frac{1}{N}\mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}_R\mathbf{U}^T$$

It can be seen that $\mathbf{\Sigma}_R$ and $\mathbf{\Sigma}_S$ have equal non-zero entries. Thus, every column of $\mathbf{V}$ is an eigenvector of $\mathbf{S}$ and every column of $\mathbf{U}$ is an eigenvector of $\mathbf{R}$. We have:

$$\begin{aligned}(\boldsymbol{v'}_i)^T &= \frac{1}{\sqrt{N}}\boldsymbol{u_i}^T\mathbf{X} \\ &= \sum_{n=1}^{N} \sigma_n(\boldsymbol{u_i}^T\mathbf{u}_n)\boldsymbol{v_n}^T \\ &= \sigma_i \boldsymbol{v_i}^T\end{aligned}$$

$$\mathbf{v}_i = \frac{\boldsymbol{v'}_i}{||\boldsymbol{v'}_i||}$$

Thus, we've have the eigenvector of $\mathbf{S}$ using that of $\mathbf{R}$. It takes $\mathcal{O}(D^2)$ complexity in power method. Here it takes $\mathcal{O}(ND)$ and is better as $N < D$.

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 0

Student Name: Sakshi
Roll Number: 180653
Date: December 19, 2020

QUESTION 2

Estimation Maximization for Poisson Mixture Model:

1. Initialize: $\boxed{\Theta = \{\pi_l, \lambda_l\}_{l=1}^{L} \quad as \quad \Theta^{(0)} \quad with \quad t = 1}$

2. Next is the Estimation step in which we will compute the conditional posterior $\mathbf{p}(\mathbf{Z}|\mathbf{K}, \Theta^{(t-1)})$ As the observations are independent and identically distributed :

$$\mathbf{p}(\mathbf{z}_n = l|\mathbf{k}_n, \Theta^{(t-1)}) \propto \mathbf{p}(\mathbf{z}_n = l, \Theta^{(t-1)})\mathbf{p}(\mathbf{k}_n|\mathbf{z}_n = l, \Theta^{t-1})$$

$$\implies \mathbf{p}(\mathbf{z}_n = l|\mathbf{k}_n, \Theta^{(t-1)}) = \pi_l^{(t-1)} \prod_{m=1}^{M} \mathbf{Poisson}(\mathbf{k}_{n,m}|\lambda_l)$$

$$= \mathbf{p}(\mathbf{z}_n = l|\mathbf{k}_n, \Theta^{(t-1)}) = \pi_l^{(t-1)} \prod_{m=1}^{M} \frac{(\lambda_l^{(t-1)})^{(\mathbf{k}_{n,m})} e^{-(\lambda_l^{(t-1)})}}{\mathbf{k}_{n,m}}$$

Hence, $\boxed{\gamma_{nl}^{(t)} = \frac{\pi_l^{(t-1)} \prod_{m=1}^{M} \mathbf{Poisson}(\mathbf{k}_{n,m}|\lambda_l)}{\sum_{i=1}^{L} \pi_i^{(t-1)} \prod_{m=1}^{M} \mathbf{Poisson}(\mathbf{k}_{n,m}|\lambda_i)}}$

3. Estimation is followed by Maximization step which maximizes the expected complete data log likelihood. Update equation is as follows:

$$\begin{aligned}
\Theta^{(t)} &= \underset{\Theta}{\text{argmax}} \mathbf{E}_{\mathbf{p}(\mathbf{Z}^{(t-1)}|\mathbf{K}, \Theta^{(t-1)})} \left[\log \mathbf{p}(\mathbf{Z}^{(t-1)}, \mathbf{K}|\Theta) \right] \\
&= \underset{\Theta}{\text{argmax}} \sum_{n=1}^{N} \mathbf{E}_{\mathbf{p}(\mathbf{Z}^{(t-1)}|\mathbf{K}, \Theta^{(t-1)})} \left[\log \mathbf{p}(\mathbf{z}_n^{(t-1)}, \mathbf{k}_n|\Theta) \right] \\
&= \underset{\Theta}{\text{argmax}} \sum_{n=1}^{N} \mathbf{E}_{\mathbf{p}(\mathbf{Z}^{(t-1)}|\mathbf{K}, \Theta^{(t-1)})} \left[\sum_{l=1}^{L} \log \mathbf{p}(\mathbf{z}_n = l|\Theta)^{\mathbf{z}_{nl}^{t-1}} \quad + \quad \log \mathbf{p}(\mathbf{k}_n|\mathbf{z}_n = l, \Theta)^{\mathbf{z}_{nl}^{(t-1)}} \right] \\
&= \underset{\Theta}{\text{argmax}} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbf{E} \left[\mathbf{z}_{nl}^{t-1} \right] \left[\log \mathbf{p}(\mathbf{z}_n = l|\Theta) + \log \mathbf{p}(\mathbf{k}_n|\mathbf{z}_n = l, \Theta) \right] \\
&= \underset{\Theta}{\text{argmax}} \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl}^{(t)} \left[\log \pi_l + \sum_{m=1}^{M} \log \mathbf{p}(\mathbf{k}_{n,m}|\mathbf{z}_n = l, \Theta) \right] \\
&= \underset{\Theta}{\text{argmax}} \sum_{n=1}^{N} \sum_{l=1}^{L} \gamma_{nl}^{(t)} \left[\log \pi_l + \left(\sum_{m=1}^{M} \mathbf{k}_{n,m} \right) \log \lambda_l - M\lambda_l \right]
\end{aligned}$$

Optimization constraints: $\sum_{l=1}^{L} \pi_l = 1$ and $\lambda_l > 0$, l = 1, 2, . . . , L. Using Lagrangian operator, we get the required updates: $\boxed{\pi_l{}^{(t)} = \frac{N_l}{N} \quad \& \quad \lambda_l{}^{(t)} = \frac{1}{MN_l} \sum_{n=1}^{N} \sum_{m=1}^{M} \gamma_{nl}{}^{(t)} k_{n,m}}$

4. $\boxed{\text{Iterate : t=t+1 until convergence}}$

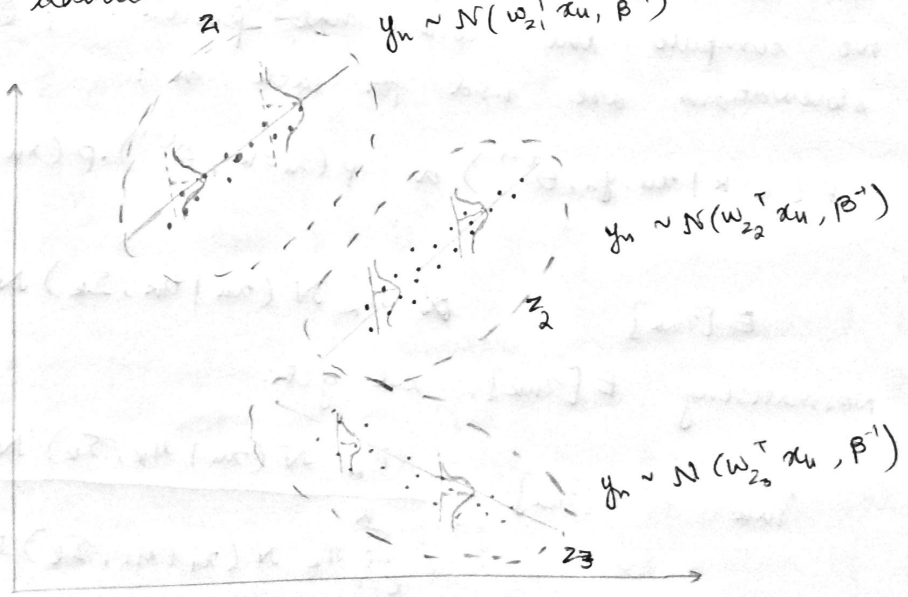
Question 3

Part 1 :

- The latent variable model for regression is doing linear regression and clustering simultaneously. We are learning K-linear regressions for each cluster. The inputs x_n are assumed to belong to one of the K-clusters. Each cluster z_n is modelled as a Gaussian distribution $N(\mu_{z_n}, \Sigma_{z_n})$. The output y_n is modelled as a cluster dependent linear transformation of input added with Gaussian noise.

$$y_n = w_{z_n}^T x_n + \epsilon \quad ; \quad \epsilon \sim N(0, \beta^{-1})$$

Thus, this model can be used for non-linearly distributed dataset. In the standard linear regression problem, a single linear transformation is used for the whole dataset. Thus we learn ~~same~~ μ & Σ for the entire dataset. Here, we first find the clusters and then learn μ_{z_k} and Σ_{z_k} for each cluster. Standard linear regression works well only for linearly ~~separable~~ distribute dataset.



- **CLL:**

$\theta = \{\pi_k, \mu_k, \Sigma_k, w_k\}_{k=1}^{K}$

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

$$= \log \prod_{n=1}^{N}\prod_{k=1}^{K} \left\{ p(x_n \mid z_n=k,\theta)\, p(y_n \mid z_n=k, x_n,\theta)\, p(z_n=k\mid\theta)\right\}^{z_{nk}}$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{ \log p(x_n \mid z_n=k,\theta) + \log p(y_n \mid z_n=k, x_n,\theta) + \log p(z_n=k\mid\theta)\right\}$$

* $\log p(x_n \mid z_n=k,\theta)$: $\mathcal{N}(x_n \mid \mu_k, \Sigma_k)$
* $\log p(y_n \mid z_n=k,\theta)$: $\mathcal{N}(y_n \mid w_{z_n}^T x_n, \beta^{-1})$
* $\log p(z_n=k)$: π_k

$$CLL(\theta) \propto \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left\{ -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x_n-\mu_k)^T\Sigma_k^{-1}(x_n-\mu_k) - \frac{\beta}{2}(y_n - w_k^T x_n)^2 + \log\pi_k\right\}$$

- **EM for Variable Regression**

- **E-Step:**

We need the expected CLL.

We compute the conditional posterior $p(z \mid x, y, \theta^{(t-1)})$. since observations are i.i.d for each n:

$$p(z_n=k \mid x_n, y_n, \theta^{t-1}) \propto p(z_n=k \mid \theta^{t-1}) \cdot p(x_n \mid z_n=k, \theta^{(t-1)}) \cdot p(y_n \mid x_n, z_n=k, \theta^{t-1})$$

$$\therefore \quad E[z_{nk}] \propto \pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\, \mathcal{N}(y_n \mid w_k^T x_n, \beta^{-1})$$

Normalizing $E[z_{nk}]$, we get:

$$\gamma_{nk} = E[z_{nk}] = \frac{\pi_k \mathcal{N}(x_n \mid \mu_k, \Sigma_k)\, \mathcal{N}(y_n \mid w_k^T x_n, \beta^{-1})}{\sum_{\ell=1}^{K} \pi_\ell \mathcal{N}(x_n \mid \mu_\ell, \Sigma_\ell)\, \mathcal{N}(y_n \mid w_\ell^T x_n, \beta^{-1})}$$

M-step: Updating θ^{t-1} by maximizing the Expected CLL.

$$\theta^t = \arg\max_{\theta} E_{P(z^{t-1} \mid X, y, \theta^{t-1})} \left[\log P(z^{(t-1)}, X, y \mid \theta) \right]$$

As we have already seen, this comes out to be:

$$\theta^t = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}^{(t-1)} \left[\log \pi_k + \log \mathcal{N}(x_n \mid \mu_k, \Sigma_k) - \frac{\beta}{2} (y_n - w_k^T x_n)^2 - \frac{1}{2} \log |\Sigma_k| \right]$$

This is to be done with respect to constraints:

$$\sum_{l=1}^{L} \pi_l = 1.$$

As seen in class, we can use lagrangian method to get:

$$\pi_k^{(t)} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk}^{(t)} = \frac{N_k}{N} \qquad N_k = \sum_{n=1}^{N} \gamma_{nk}^{(t)}$$

$$\Sigma_k^{(t)} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}^{(t)} (x_n - \mu_k^t)(x_n - \mu_k^t)^T$$

$$\mu_k^t = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_{nk}^{(t)} x_n \; ; \; \hat{w}_k = \left(\sum_n \gamma_{nk} x_n x_n^T \right)^{-1} \left(\sum_n \gamma_{nk} y_n x_n \right)$$

Thus, the overall algorithm is:

1. Initialize $\theta = \theta^0$; $t = 1$
2. Perform the E-step
3. Perfor the M-step

] as described above

4. $t = t+1$; Go to step 2 if not converged.

- Update Equation:

Our update equation for w_k is similar to the standard linear regression one only with extra γ_{nk}. Here, only those points are considered for k^{th} cluster which have $\gamma_{nk} = 1$. $\gamma_{nk} = 0$ ~~to~~ to $\gamma_{nk} = 1$ determines the contribution of a point in a cluster. Thus, linear regression happens with expected value of points determining how much a point contributes in a cluster.

- **ALT- OPT Algorithm:**

1. Initialize $\theta = \{\mu_k, \Sigma_k, w_k\}_{k=1}^{K}$; Set $t=1$;

2. Here we'll take point estimate of z_n instead of $E[z_n]$. Let's compute the most probable value of z_n as:

$$\hat{z_n} = \arg\max_k \pi_k \mathcal{N}(x_n \mid \mu_k^{(t-1)}, \Sigma_k^{(t-1)}) \mathcal{N}(y_n \mid (w_k^{T})^{t-1} x_n, \beta^{-1})$$

Given all $\pi_k = \frac{1}{K}$, we can neglect it.

3. We can solve the MLE problem for θ using z_n. Here, we don't need expected value as we have the actual z_ns. So we can just replace γ_{nk}s by z_{nk}s, got in the ALT- OPT method.

Thus, we get:

$$\mu_k^{(t)} = \frac{1}{N} \sum_{n=1}^{N} z_{nk} \qquad N_k = \sum_{n=1}^{N} z_{nk}$$

$$\Sigma_k^{(t)} = \frac{1}{N_k} \sum_{n=1}^{N} z_{nk} (x_n - \hat{\mu_k})(x_n - \hat{\mu_k})^T$$

$$\hat{w_k}^{(t)} = \left(\sum_{n: z_{nk}=1} x_n x_n^T \right) \left(\sum_{n: z_{nk}=1} y_n x_n \right)$$

4. $t = t+1$; Go to Step 2 if not converged.