Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

QUESTION 1

Student Name: Machine Learner
Roll Number: 17001
Date: December 19, 2020

An eigen vector v of $\frac{1}{N}XX^T$ or XX^T, satisfies a relation $XX^Tv = \lambda v$.
Multiplying both sides by X^T and using associativity of matrix multiplixation,

$$\frac{1}{N}X^T(XX^T)v = \frac{1}{N}X^TX(X^Tv) = \frac{1}{N}X^T\lambda v = \frac{1}{N}\lambda(X^Tv)$$

Thus,

$$\frac{1}{N}X^TX(X^Tv) = \frac{1}{N}\lambda(X^Tv)$$

Hence, we see that for each eigen vector (with a non-zero eigen-value) of $\frac{1}{N}XX^T$, we can derive an eigen vector for $\frac{1}{N}X^TX$ by pre-multiplying it by X^T.
Now, to obtain those eigen-vectors of $\frac{1}{N}X^TX$ for which the eigen value is 0, we do the following. Let's assume u is an eigen vector of X^TX with 0 eigenvalue.

$$X^TX(u) = 0 \implies u^TX^TXu = 0 \implies ||Xu||^2 = 0$$

Hence, all the vectors in the null-space of X are the eigen vectors of $\frac{1}{N}X^TX$ with 0 eigen value. The advantage of using this method is that we need to compute eigen values for a NxN matrix if we use XX^T, and hence require a computation time of $O(N^3)$, as opposed to the $O(D^3)time$ for calculating eigen vectors of the DXD matrix X^TX. Since it is given in the question that $D > N$, we achieve a computational advantage by doing so.

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Machine Learner
Roll Number: 17001
Date: December 19, 2020

QUESTION 2

The complete log data likelihood for the model can be written as:

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

E-step:
We compute the posterior conditional distribution of z as:

$$\begin{aligned}p(z_{nl}=1|k_n,\theta) &\propto p(z_{nl}=1|\theta).p(k_n|z_{nl}=1,\theta)\\ &=\pi_l.\prod_{m=1}^{M}\frac{\lambda_l^{k_{n,m}}}{e^{\lambda_l}.k_{n,m}!}\end{aligned}$$

Here θ is used to denote the collective set of parameters. In order to compute $\mathbb{E}[z_nl]$, we do:

$$\mathbb{E}[z_{nl}^{(t)}]=\gamma_{nl}^{(t)}=\frac{\pi_l^{(t-1)}\cdot\prod_{m=1}^{M}\frac{{\lambda_l^{(t-1)}}^{k_{n,m}}}{e^{\lambda_l^{(t-1)}}.k_{n,m}!}}{\Sigma_{o=1}^{L}\pi_o^{(t-1)}\cdot\prod_{m=1}^{M}\frac{{\lambda_o^{(t-1)}}^{k_{n,m}}}{e^{\lambda_o^{(t-1)}}.k_{n,m}!}}$$

M-step:
The expression for maximizing expected CLL can be written as:

$$\underset{\pi,\lambda}{\arg\max}\quad\Sigma_{n=1}^{N}\Sigma_{l=1}^{L}\mathbb{E}[z_{nl}]\left[log(\pi_l)+\Sigma_{m=1}^{M}log(\frac{\lambda_l^{k_{n,m}}}{e^{\lambda_l}.k_{n,m}!})\right]$$

Differentiating wrt π_l, we get

$$\begin{aligned}\Sigma_{n=1}^{N}\mathbb{E}[z_{nl}]\left[\frac{1}{\pi_l}\right]&=0\\ \implies \pi_l=\Sigma_{n=1}^{N}\mathbb{E}[z_{nl}]&\quad\forall l\in[1,L]\end{aligned}$$

Differentiating wrt λ_l, we get

$$\begin{aligned}\Sigma_{n=1}^{N}\mathbb{E}[z_{nl}]\Sigma_{m=1}^{M}\left[\frac{k_{n,m}}{\lambda_l}-1\right]&=0\\ \implies \lambda_l=\frac{\Sigma_{n=1}^{N}\mathbb{E}[z_{nl}]\Sigma_{m=1}^{M}k_{n,m}}{M\Sigma_{n=1}^{N}\mathbb{E}[z_{nl}]}&\quad\forall l\in[1,L]\end{aligned}$$

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 1

Student Name: Machine Learner
Roll Number: 17001
Date: December 19, 2020

QUESTION 3

In a standard linear regression model, x_n and y_n are provided to us, and we just learn a global weight vector w, which translates the inputs to the appropriate y_n.
Here, the generative story is as follows. First, we generate a latent variable z_n belonging to one of the K classes, derived from the multinoulli distribution of z. Next, we sample an observation x_n from a gaussian, whose mean and variance are fixed but dependent on the class k that z_n belongs to. Once we have the value of x_n, we map it to another gaussian distribution whose mean is dependent on x_n, with its variance being a fixed quantity β^{-1}. We sample y_n from this gaussian distribution, which is the output for x_n.
The advantage of using this is that we are able to learn K different distributions for input data, and based on which distribution an input came from, we associate another distribution for y_n, from which we sample the final output of our model.
Essentially, we learn K different weight vectors for our linear regression model, and depending on the input, we decide which weight vector to choose.
After we generate z_n and fixing the distribution of x_n, we basically perform a probabilistic PCA for finding y_n.
Thus, y_n can be seen as a low-dimensional representation of the input x_n, which is exactly what we want to learn in the regression model.

EM Algorithm:

1. Initialize $\Theta = \{\pi_k, \mu_k, \Sigma_k, \mathbf{w}_k\}_{k=1}^{K}$ as $\Theta^{(0)}$, set $t = 1$

2. **E step:** Compute the following for all z_n.
Conditional posterior distr of Z:

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

This is the unnormalized version of Conditional posterior distr of Z. Upon calculation of $\mathbb{E}[z_{nk}]$, we get

$$\mathbb{E}[z_{nk}^{(t)}] = \gamma_{nk}^{(t)} = \frac{\pi_k^{(t-1)}.\mathcal{N}(\mu_k^{(t-1)}, \Sigma_k^{(t-1)}).\mathcal{N}(\mathbf{w_k^{(t-1)T} x_n}, \beta^{-1})}{\Sigma_{l=1}^{K}\pi_l^{(t-1)}.\mathcal{N}(\mu_l^{(t-1)}, \Sigma_l^{(t-1)}).\mathcal{N}(\mathbf{w_l^{(t-1)T} x_n}, \beta^{-1})}$$

3. **M step:** Compute **CLL** and maximize its expectation:

$$\begin{aligned} p(\mathbf{X}, \mathbf{Y}, \mathbf{Z}|\theta) &= \prod_{n=1}^{N} p(z_n|\theta).p(x_n|z_n, \theta).p(y_n|z_n, x_n, \theta) \\ &= \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\pi_k.\mathcal{N}(x_n|\mu_k, \Sigma_k).\mathcal{N}(y_n|\mathbf{w_k^T x_n}, \beta^{-1})\right]^{z_{nk}} \end{aligned}$$

Therefore, CLL is given as:

$$log(p(\mathbf{X},\mathbf{Y},\mathbf{Z}|\theta)) = \Sigma_{n=1}^{N}\Sigma_{k=1}^{K} z_{nk}\left[log(\pi_k) + log(\mathcal{N}(x_n|\mu_k,\Sigma_k)) + log(\mathcal{N}(y_n|\mathbf{w_k^T x_n},\beta^{-1}))\right]$$

In this step, we maximize the expected CLL wrt Θ, as:

$$\arg\max_{\Theta}\ \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}\mathbb{E}[z_{nk}]\left[log(\pi_k) + log(\mathcal{N}(x_n|\mu_k,\Sigma_k)) + log(\mathcal{N}(y_n|\mathbf{w_k^T x_n},\beta^{-1}))\right]$$

On simplifying, we get

$$\arg\max_{\Theta}\ \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}\mathbb{E}[z_{nk}]\left[log(\pi_k) + log(\mathcal{N}(x_n|\mu_k,\Sigma_k)) - \frac{(y_n - \mathbf{w_k^T x_n})^2\beta^2}{2}\right]$$

Differentiating wrt μ_k, we get

$$\mu_k^{(t)} = \frac{\Sigma_{n=1}^{N}\gamma_{nk}^{(t)}x_n}{\Sigma_{n=1}^{N}\gamma_{nk}^{(t)}}$$

Differentiating wrt Σ_k, we get

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

Differentiating wrt π_k, we get

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

Differentiating wrt $\mathbf{w}_k$, we get

$$\Sigma_{n=1}^{N}\mathbb{E}[z_{nk}]\left(x_n(y_n - w_k^T x_n)\right) = 0$$
$$\implies w_k = (\Sigma_{n=1}^{N}\mathbb{E}[z_{nk}]x_n x_n^T)^{-1}\Sigma_{n=1}^{N}(\mathbb{E}[z_{nk}]y_n)x_n$$

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

Intuitive Sense: The update is very similar to the closed form solution obtained in normal linear regression $W = (X^TX)^{-1}X^TY$

ALT-OPT Algorithm:

1. Initialize $\Theta = \{\pi_k, \mu_k, \Sigma_k, \mathbf{w}_k\}_{k=1}^{K}$ as $\Theta^{(0)}$, set $t = 1$
2. **Conditional posterior distr of Z:**

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

For each n, compute the best guess of z_n as:

$$\hat{z_n} = \arg\max_{k=1,2....K} p(z_n = k|x_n, y_n, \theta) \quad (1)$$

3. **S**olve MLE problem for θ using $\hat{z_n}$ in the last step

$$\hat{\Theta} = \arg\max_{\Theta} \ \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}\hat{z_{nk}} \left[log(\pi_k) + log(\mathcal{N}(x_n|\mu_k,\Sigma_k)) + log(\mathcal{N}(y_n|\mathbf{w_k^T x_n},\beta^{-1}))\right]$$

On simplifying, we get

$$\arg\max_{\Theta} \ \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}\hat{z_{nk}} \left[log(\frac{1}{k}) + log(\mathcal{N}(x_n|\mu_k,\Sigma_k)) - \frac{(y_n - \mathbf{w_k^T x_n})^2\beta^2}{2}\right]$$

Differentiating wrt μ_k, we get

$$\mu_k^{(t)} = \frac{\Sigma_{n=1}^{N}\hat{z_{nk}}^{(t)}x_n}{\Sigma_{n=1}^{N}\hat{z_{nk}}^{(t)}}$$

Differentiating wrt Σ_k, we get

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

Differentiating wrt $\mathbf{w}_k$, we get

$$\Sigma_{n=1}^{N}\hat{z_{nk}}\left(x_n(y_n - w_k^T x_n)\right) = 0$$
$$\implies w_k = (\Sigma_{n=1}^{N}\hat{z_{nk}}x_n x_n^T)^{-1}\Sigma_{n=1}^{N}(\hat{z_{nk}}y_n)x_n$$

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

Part 2:

EM Algorithm:

1. Initialize $\Theta = \{\eta_k, \mathbf{w}_k\}_{k=1}^{K}$ as $\Theta^{(0)}$, set $t = 1$
2. **E step:** Compute the following for all z_n.
Conditional posterior distr of Z:

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

This is the unnormalized version of Conditional posterior distr of Z. Upon calculation of $\mathbb{E}[z_{nk}]$, we get

$$\mathbb{E}[z_{nk}^{(t)}] = \gamma_{nk}^{(t)} = \frac{\pi_k(x_n)\mathcal{N}(\mathbf{w_k^T x_n},\beta^{-1})}{\Sigma_{l=1}^{K}\pi_l(x_n)\mathcal{N}(\mathbf{w_l^T x_n},\beta^{-1})}$$

3. **M step:** Compute **CLL** and maximize its expectation:

$$p(\mathbf{X},\mathbf{Y},\mathbf{Z}|\theta) == \prod_{n=1}^{N}\prod_{k=1}^{K}\left[\pi_k(x_n).\mathcal{N}(y_n|\mathbf{w_k^T x_n},\beta^{-1})\right]^{z_{nk}}$$

Therefore, CLL is given as:

$$log(p(\mathbf{X},\mathbf{Y},\mathbf{Z}|\theta)) = \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}z_{nk}\left[log(\pi_k(x_n)) + log(\mathcal{N}(y_n|\mathbf{w_k^T x_n},\beta^{-1}))\right]$$

In this step, we maximize the expected CLL wrt Θ, as:

$$\arg\max_{\Theta} \ \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}\mathbb{E}[z_{nk}]\left[log(\pi_k(x_n)) + log(\mathcal{N}(y_n|\mathbf{w_k^T x_n}, \beta^{-1}))\right]$$

On simplifying, we get

$$\arg\max_{\Theta} \ \Sigma_{n=1}^{N}\Sigma_{k=1}^{K}\mathbb{E}[z_{nk}]\left[log(\pi_k(x_n)) - \frac{(y_n - \mathbf{w_k^T x_n})^2\beta^2}{2}\right]$$

Differentiating wrt η_k, we get

$$\frac{1}{\pi_k(x_n)}.\frac{d(\pi_k(x_n))}{d\eta_k} = 0$$

Differentiating wrt $\mathbf{w}_k$, we get

$$\Sigma_{n=1}^{N}\mathbb{E}[z_{nk}]\left(x_n(y_n - w_k^T x_n)\right) = 0$$
$$\implies w_k = (\Sigma_{n=1}^{N}\mathbb{E}[z_{nk}]x_n x_n^T)^{-1}\Sigma_{n=1}^{N}(\mathbb{E}[z_{nk}]y_n)x_n$$

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