Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 1

Student Name: Utkarsh Gupta
Roll Number: 180836
Date: December 19, 2020

Let the matrix $\mathbf{R} = \frac{1}{N}\mathbf{X}\mathbf{X}^T$. We are given an eigenvector $\mathbf{v} \in \mathbb{R}^N$ of $\mathbf{R}$ (eigenvalue λ_v) and we have to find eigenvector $\mathbf{u} \in \mathbb{R}^D$ of $\mathbf{S}$ (eigenvalue λ_u). From the definition of eigenvectors, we know:

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

$$\implies \frac{1}{N}\left(\mathbf{X}^T\mathbf{X}\right)\mathbf{X}^T\mathbf{v} = \lambda_v\left(\mathbf{X}^T\mathbf{v}\right)$$

$$\implies \frac{1}{N}\mathbf{S}\left(\mathbf{X}^T\mathbf{v}\right) = \lambda_u\left(\mathbf{X}^T\mathbf{v}\right)$$

Thus we have,

$$\mathbf{u} = \mathbf{X}^T\mathbf{v} \;\; \& \;\; \lambda_u = \lambda_v$$

We know $D > N$ and $rank(\mathbf{S}) = rank(\mathbf{R}) = min(N, D) = N$. $\mathbf{S}$ is a DxD matrix, which means that it has D eigenvectors, N of which can be found using the relation found above and the remaining $D - N$ vectors will be $\mathbf{0}$.
The main advantage of using this method is the speed that we gain. Eigendecomposition of a DxD (original $\mathbf{S}$) takes $O(D^3)$ time. Using this method, we can do eigendecomposition of $\mathbf{R}$ in $O(N^3)$ time and then multiply them with $\mathbf{X}^T$ in a total time of $O(N^2D)$. Thus, the total time complexity of this approach is $O(N^3) + O(N^2D) = O(N^2D)$ which is significantly less than the original $O(D^3)$. Hence, this approach will be much faster when D is much larger than N.

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 3

QUESTION 2

Student Name: Utkarsh Gupta
Roll Number: 180836
Date: December 19, 2020

We are given $\mathbf{K}$ an N x M matrix that contains the per minute server hit data and L (number of clusters) which is a hyper-parameter. $\boldsymbol{\lambda}$ is an L x 1 vector, l^{th} entry of which is the Poisson parameter of the l^{th} cluster.
To find the CLL we need:

$$p(\mathbf{K}, \mathbf{Z}|\boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{n=1}^{N} \sum_{l=1}^{L} \left[z_{nl} \text{ x } p(z_n = l) \prod_{m=1}^{M} \text{Poisson}(k_{nm}|\lambda_l) \right]$$

$\text{Poisson}(k|\lambda_l) = \frac{\lambda^k}{k!} e^{-\lambda}$ and $p(z_n = l) = \pi_l$. Since $\mathbf{z}_n$ is one-hot vector, only one of the z_{nl} is 1.

$$CLL(\mathbf{K}, \mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \sum_{l=1}^{L} z_{nl} \left[log\ \pi_l + \sum_{m=1}^{M} (k_{nm} log\ \lambda_l - \lambda_l - log\ k_{mn}!) \right]$$

For solving, we can ignore the objective-independent constants:

$$CLL(\mathbf{K}, \mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \sum_{l=1}^{L} z_{nl} \left[log\ \pi_l + \sum_{m=1}^{M} (k_{nm} log\ \lambda_l - \lambda_l) \right] \tag{1}$$

E step:

$$\mathbb{E}[z_{nl}] = 0 \times p(z_{nl} = 0 \mid k_n, \boldsymbol{\pi}, \boldsymbol{\lambda}) + 1 \times p(z_{nl} = 1 \mid k_n, \boldsymbol{\pi}, \boldsymbol{\lambda})$$

$$\mathbb{E}[z_{nl}] = p(z_{nl} = 1 \mid k_n, \boldsymbol{\pi}, \boldsymbol{\lambda}) \propto p(z_n = l \mid \boldsymbol{\pi}) \times p(k_n \mid z_n = l, \boldsymbol{\lambda})$$

$$\implies \mathbb{E}[z_{nl}] \propto \pi_l \prod_{m=1}^{M} \frac{\lambda_l^{k_{nm}}}{k_{nm}!} e^{-\lambda_l}$$

$$\therefore \mathbb{E}[z_{nl}] = \frac{\pi_l \prod_{m=1}^{M} \frac{\lambda_l^{k_{nm}}}{k_{nm}!} e^{-\lambda_l}}{\sum_{l=1}^{L} \pi_l \prod_{m=1}^{M} \frac{\lambda_l^{k_{nm}}}{k_{nm}!} e^{-\lambda_l}} \tag{2}$$

M step:

$$\arg \max_{\boldsymbol{\pi} \geq 0, \boldsymbol{\lambda} > 0} \mathbb{E}[CLL(\mathbf{K}, \mathbf{Z} \mid \boldsymbol{\pi}, \boldsymbol{\lambda})] = \arg \max_{\boldsymbol{\pi} \geq 0, \boldsymbol{\lambda} > 0} \sum_{n=1}^{N} \sum_{l=1}^{L} \mathbb{E}[z_{nl}] \left[log\ \pi_l + \sum_{m=1}^{M} (k_{nm} log\ \lambda_l - \lambda_l) \right]$$

given $\sum_{l=1}^{L} \lambda_l = 1$. Now, the Lagrangian:

$$\arg\max_{\boldsymbol{\pi},\boldsymbol{\lambda}} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0,\theta} \mathcal{L}(\boldsymbol{\pi},\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\theta) = \mathbb{E}\left[CLL(\mathbf{K},\mathbf{Z} \mid \boldsymbol{\pi},\boldsymbol{\lambda})\right] + \sum_{l=1}^{L} \alpha_l \pi_l + \sum_{l=1}^{L} \beta_l \lambda_l - \theta(1 - \sum_{l=1}^{L} \pi_l)$$

Taking derivatives, we obtain primal variables:

$$\lambda_i = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}\left[z_{ni}\right] k_{nm}}{-\beta_i + M \sum_{n=1}^{N} \mathbb{E}\left[z_{ni}\right]}$$

$$\pi_i = -\frac{\sum_{n=1}^{N} \mathbb{E}\left[z_{ni}\right]}{\alpha_i + \theta}$$

Substituting and simplifying the primal variables, the problem reduces to:

$$\arg \min_{\boldsymbol{\alpha},\boldsymbol{\beta}\geq 0,\theta} \mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\theta) = -\sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}\left[z_{nl}\right] \left[log(-\alpha_l - \theta) + \sum_{m=1}^{M} k_{nm} log\left(-\beta_l + M \sum_{n=1}^{N} \mathbb{E}\left[z_{nl}\right]\right)\right] - \theta$$

Taking derivatives with respect to $\boldsymbol{\alpha}$ & $\boldsymbol{\beta}$, we get:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} > 0 \quad \& \quad \frac{\partial \mathcal{L}}{\partial \beta_i} > 0$$

Thus, $\mathcal{L}$ monotonically increases with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. So, $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\beta}} = \mathbf{0}$. Setting the derivative with respect to θ to 0, we get:

$$\hat{\theta} = -\sum_{n=1}^{N} \sum_{m=1}^{M} \mathbb{E}\left[z_{nl}\right] = -N$$

Finally, substituting them we get:

$$\pi_i = \frac{N_i}{N} \quad \& \quad \lambda_i = \frac{1}{N_i} \sum_{n=1}^{N} \mathbb{E}\left[z_{ni}\right] \mathbb{E}\left[k_n\right] \tag{3}$$

where $N_i = \sum_{n=1}^{N} \mathbb{E}\left[z_{ni}\right]$ is the expected size of the i^{th} cluster and $\mathbb{E}\left[k_n\right] = \frac{1}{M} \sum_{m=1}^{M} k_{nm}$ is the expected number of hits on the n^{th} server.

Introduction to ML (CS771), Autumn 2020
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Utkarsh Gupta
Roll Number: 180836
Date: December 19, 2020

QUESTION 3

The generative story for each $(\mathbf{x}_n, y_n)$ is:
1. Generate $z_n \sim \text{MULTINOULLI}(\pi_1, \pi_2, ..., \pi_K)$
2. Generate $\mathbf{x}_n \sim \mathcal{N}(\mu_{z_n}, \Sigma_{z_n})$
3. Finally generate $y_n \sim \mathcal{N}(\mathbf{w}_{z_n}^T \mathbf{x}_n, \beta^{-1})$

$\therefore$ This model behaves like an ensemble of K probabilistic linear regressors:

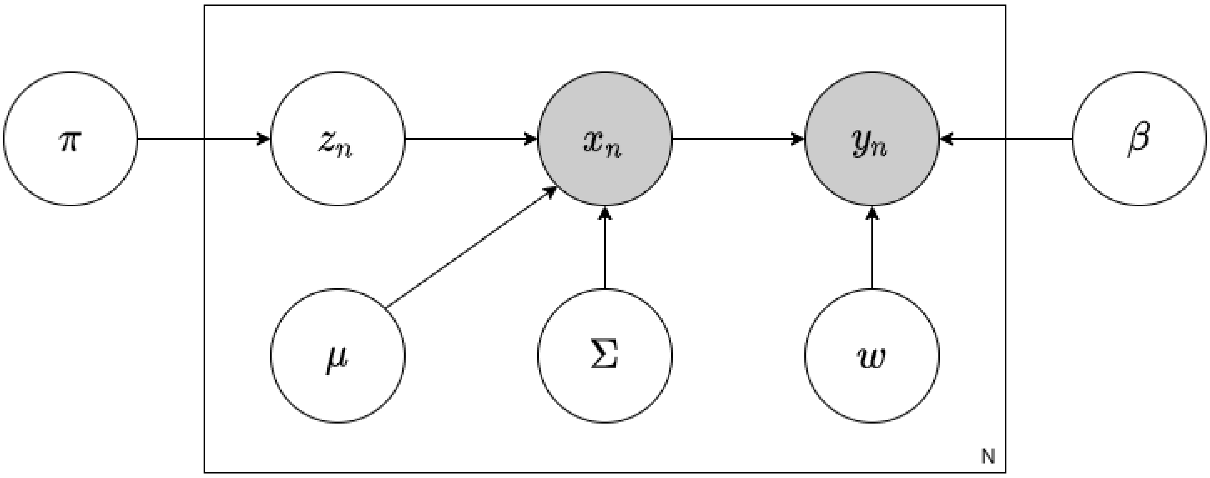

Figure 1: Pictorial Representation of the Generative Story

Part-1:

Let $\mathbf{\Theta} = \{\pi_k, \mu_k, \Sigma_k, \mathbf{w}_k\}_{k=1}^K$

The CLL:

$$\log p(\mathbf{X}, \mathbf{y}, \mathbf{Z}|\mathbf{\Theta}) = \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \beta) + \log p(\mathbf{X}|\mathbf{Z}, \mu, \Sigma) + \log p(\mathbf{Z}|\boldsymbol{\pi})$$

Assuming iid variables and substituting their expressions in the CLL we get:

$$CLL(\mathbf{X}, \mathbf{y}, \mathbf{Z}, \mathbf{\Theta}) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[-\frac{\beta}{2}(\mathbf{y}_n - \mathbf{w}_k^T \mathbf{x}_n)^2 - \frac{1}{2} \log |\mathbf{\Sigma}_k| - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_k) + \log \pi_k \right]$$

Moving on to the EM algorithm:

E-step:

$$\mathbb{E}[z_{nk}] \propto p(z_{nk} = 1|\mathbf{x}_n, \mathbf{y}_n, \boldsymbol{\Theta}) = p(z_n = k|\boldsymbol{\pi})p(\mathbf{x}_n|z_n = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)p(\mathbf{y}_n|\mathbf{x}_n, z_{nk} = 1)$$

$$\therefore \mathbb{E}[z_{nk}] \propto \pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)N(\mathbf{y}_n|\mathbf{w}_k^T\mathbf{x}_n, \beta^{-1})$$

$$\therefore \mathbb{E}[z_{nk}] = \frac{\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)N(\mathbf{y}_n|\mathbf{w}_k^T\mathbf{x}_n, \beta^{-1})}{\sum_{k=1}^{K}\pi_k N(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)N(\mathbf{y}_n|\mathbf{w}_k^T\mathbf{x}_n, \beta^{-1})}$$

M-step:

The objective function is:

$$\boldsymbol{\Theta}_{opt} = \underset{\boldsymbol{\pi}\geq 0}{\operatorname{argmax}} = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left[-\frac{\beta}{2}(\mathbf{y}_n - \mathbf{w}_k^T\mathbf{x}_n)^2 - \frac{1}{2}(\mathbf{x}_n - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_k) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| + \log\pi_k\right]$$

with the constraint $\sum_{k=1}^{K}\pi_k = 1$.

In order to deal with the constraint, we can construct a Lagrangian:

$$\mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \theta) = \mathbb{E}[CLL(\mathbf{X}, \mathbf{y}, \mathbf{Z}, \boldsymbol{\Theta})] + \sum_{k=1}^{K}\alpha_k\pi_k - \theta\left(1 - \sum_{k=1}^{K}\pi_k\right)$$

$$\boldsymbol{\Theta}_{opt} = \operatorname{argmax}\ \min_{\boldsymbol{\alpha}\geq 0}\ \mathcal{L}(\boldsymbol{\Theta}, \boldsymbol{\alpha}, \theta)$$

On differentiating the Lagrangian w.r.t primal variables we get:

$$\frac{\partial\mathcal{L}}{\partial\mathbf{w}_i} = \beta\mathbf{X}^T diag(\mathbb{E}[Z_i])(\mathbf{y} - \mathbf{X}\mathbf{w}_i) = 0$$

$$\implies \mathbf{w}_i = [\mathbf{X}^T diag(\mathbb{E}[Z_i])\mathbf{X}]^{-1}\mathbf{X}^T diag(\mathbb{E}[Z_i])\mathbf{y}$$

where diag($\mathbb{E}[\mathbf{Z}_i]$) is a N x N diagonal matrix with $\mathbb{E}[z_{ji}]$ as the j^{th} diagonal entry.
The expressions obtained for μ_i and σ_i are:

$$\boldsymbol{\mu}_i = \frac{1}{N_i}\sum_{n=1}^{N}\mathbb{E}[z_{ni}]\mathbf{x}_n$$

$$\boldsymbol{\Sigma}_i = \frac{1}{N_i}\sum_{n=1}^{N}\mathbb{E}[z_{ni}](\mathbf{x}_n - \boldsymbol{\mu}_i)(\mathbf{x}_n - \boldsymbol{\mu}_i)^T$$

$$\frac{\partial\mathcal{L}}{\partial\pi_i} = \sum_{n=1}^{N}\frac{\mathbb{E}[z_{ni}]}{\pi_i} + \alpha_i + \theta = 0 \implies \pi_i = \frac{N_i}{-\alpha_i - \theta}$$

where $N_i = \sum_{n=1}^{N}\mathbb{E}[z_{ni}]$
Formulating the dual and removing unwanted constants we have:

$$\boldsymbol{\alpha}_{opt}, \theta_{opt} = \underset{\boldsymbol{\alpha}\geq 0}{\operatorname{argmin}}\sum_{n=1}^{N}\sum_{k=1}^{K} -\mathbb{E}[z_{nk}]\log(-\alpha_k - \theta) - \theta$$

Taking its derivative w.r.t α_i we have:

$$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \sum_{n=1}^{N} \frac{\mathbb{E}[z_{ni}]}{-\alpha_i - \theta} \geq 0$$

$$\frac{\partial^2 \mathcal{L}}{\partial \alpha_i^2} = \sum_{n=1}^{N} \frac{\mathbb{E}[z_{ni}]}{(\alpha_i + \theta)^2} \geq 0$$

Thus, $\mathcal{L}$ increases monotonically in $\boldsymbol{\alpha}$, $\boldsymbol{\alpha}_{opt} = 0$

$$\therefore \theta = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}[z_{nk}] = -N$$

$$\implies \pi_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}[z_{ni}]$$

The EM algorithm can be summarised as follows:

1. Initialize $\boldsymbol{\Theta} = \hat{\boldsymbol{\Theta}}$
2. E-step: Calculate the $\mathbb{E}[z_{nk}]$ {for k = 1 to K, n=1 to N}
3. M-step: Estimate $\boldsymbol{\Theta}$'s parameters using the expressions given for $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \pi_i$ and $\mathbf{w}_i$ above {for k=1 to K}
4. Go to step-2 if not converged

The expression obtained for weights makes intuitive sense because it is analogous to the weight vector in an importance-weighted linear regression problem. Moreover, if we consider each class as the output of a single regression problem, these expressions start to make intuitive sense.

Moving on, for the ALT-OPT algo, we have $\pi_k = \frac{1}{K}$, so $\boldsymbol{\Theta} = \{\mu_k, \Sigma_k, w_k\}_{k=1}^{K}$

Step-1: MLE estimate of $\mathbf{z}_n$:

$$\hat{\mathbf{z}}_n = \underset{k}{\operatorname{argmax}}\, p(z_n = k | \mathbf{X}, \mathbf{y}, \boldsymbol{\Theta}) \propto p(z_n = k | \boldsymbol{\pi}) p(\mathbf{x}_n, \mathbf{y}_n | z_n = k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Transforming it to a minimisation problem and removing constants:

$$\hat{\mathbf{z}}_n = \underset{k}{\operatorname{argmin}}[(\mathbf{x}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k) + \log |\boldsymbol{\Sigma}_k| + \frac{\beta}{2}(\mathbf{y}_n - \mathbf{w}_k^T \mathbf{x}_n)^2]$$

$E[z_{nk}]$ can be replaced by $\hat{z}_{nk}$:

$$\therefore \boldsymbol{\mu}_i = \frac{1}{N_i} \sum_{n=1}^{N} \hat{z}_{ni} \mathbf{x}_n$$

$$\mathbf{\Sigma}_i = \frac{1}{N_i}\sum_{n=1}^{N} \hat{z}_{ni}(\mathbf{x}_n - \boldsymbol{\mu}_i)(\mathbf{x}_n - \boldsymbol{\mu}_i)^T$$

$$\mathbf{w}_i = [\mathbf{X}^T diag(\hat{\mathbf{Z}}_i)\mathbf{X}]^{-1}\mathbf{X}^T diag(\hat{\mathbf{Z}}_i)\mathbf{y}$$

where $\hat{z}_{nk} = 1$ if $\hat{z}_n = k$ else 0.
ALT-OPT can be summarised as:

1. Initialize $\mathbf{\Theta} = \hat{\mathbf{\Theta}}$
2. Estimate $\mathbf{z}_i$ for $i = 1$ to N by solving the minimisation expression shown above.
3. Estimate $\mathbf{\Theta}$ using the expressions given above {for k = 1 to K}.
4. Go to step-2 if not converged.

Part-2:
Let the params be $\mathbf{\Theta} = \{\boldsymbol{\eta}_k, \mathbf{w}_k\}_{k=1}^{K}$:
Now for the EM, expression for CLL is:

$$\log p(\mathbf{y}, \mathbf{Z}|\mathbf{X}, \mathbf{\Theta}) = \log p(\mathbf{y}|\mathbf{X}, \mathbf{W}, \beta) + \log p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\eta})$$

$$CLL(\mathbf{y}, \mathbf{Z}, \mathbf{X}, \mathbf{\Theta}) = \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left[-\frac{\beta}{2}(y_n - \mathbf{w}_K^T\mathbf{x}_n)^2 + \boldsymbol{\eta}^T\mathbf{x}_n - \log\sum_{j=1}^{K}\exp(\boldsymbol{\eta}_j^T\mathbf{x}_n)\right]$$

For the expectation step:

$$\mathbb{E}[z_{nk}] = \frac{\sum_{k=1}^{K} z_{nk}p(z_n = k|\mathbf{X}, \mathbf{\Theta})}{\sum_{k=1}^{K} p(z_n = k|\mathbf{X}, \mathbf{\Theta})} = \frac{\sum_{k=1}^{K} z_{nk}\pi_k(\mathbf{x}_n)}{\sum_{k=1}^{K}\pi_k(\mathbf{x}_n)}$$

$$\implies \mathbb{E}[z_{nk}] = \frac{\exp(\boldsymbol{\eta}_k^T\mathbf{x}_n)}{\sum_{k=1}^{K}\exp(\boldsymbol{\eta}_k^T\mathbf{x}_n)}$$

For the maximization step, the objective is:

$$\mathbf{\Theta}_{opt} = \mathbb{E}[CLL(\mathbf{y}, \mathbf{Z}, \mathbf{X}, \mathbf{\Theta})] = \sum_{n=1}^{N}\sum_{k=1}^{K}\mathbb{E}[z_{nk}]\left[-\frac{\beta}{2}(y_n - \mathbf{w}_K^T\mathbf{x}_n)^2 + \boldsymbol{\eta}^T\mathbf{x}_n - \log\sum_{j=1}^{K}\exp(\boldsymbol{\eta}_j^T\mathbf{x}_n)\right]$$

This gives us an unconstrained problem, which can be solved by differentiating:

$$\mathbf{w}_i = [\mathbf{X}^T diag(\hat{\mathbf{Z}}_i)\mathbf{X}]^{-1}\mathbf{X}^T diag(\hat{\mathbf{Z}}_i)\mathbf{y}$$

$$\frac{\partial\mathcal{L}}{\partial\boldsymbol{\eta}_i} = \sum_{n=1}^{N}\mathbb{E}[z_{ni}]\left[\mathbf{x}_n - \frac{\mathbf{x}_n\exp(\boldsymbol{\eta}_i^T\mathbf{x}_n)}{\sum_{j=1}^{K}\exp(\boldsymbol{\eta}_j^T\mathbf{x}_n)}\right] = 0$$

$$\therefore \sum_{n=1}^{N}\frac{\mathbb{E}[z_{ni}]\exp(\boldsymbol{\eta}_i^T\mathbf{x}_n)}{\sum_{j=1}^{K}\exp(\boldsymbol{\eta}_j^T\mathbf{x}_n)} = \sum_{n=1}^{N}\mathbb{E}[z_{ni}]\mathbf{x}_n$$

This gives us a system of K non-linear equations for each (i=1,...,K), solving which will give the required $\boldsymbol{\eta}_i$. It is not possible to get a closed form expression for $\boldsymbol{\eta}_i$ as they are dependent on each other's value.