Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 1

$$f(\alpha+\delta\mathbf{e}_n) = (\alpha+\delta\mathbf{e}_n)^T\mathbf{1} - \frac{1}{2}(\alpha+\delta\mathbf{e}_n)^T\mathbf{G}(\alpha+\delta\mathbf{e}_n)^T$$

$$= \alpha^T\mathbf{1} + \delta\mathbf{e}_n^T\mathbf{1} - \frac{1}{2}(\alpha^T\mathbf{G}\alpha + \delta\alpha^T\mathbf{G}\mathbf{e}_n + \delta\mathbf{e}_n^T\mathbf{G}\alpha + \delta^2\mathbf{e}_n^T\mathbf{G}\mathbf{e}_n)$$

$$\delta_* = \arg\max_\delta f(\alpha+\delta\mathbf{e}_n)$$

$$\nabla_\delta f(\alpha+\delta\mathbf{e}_n) = 0$$

$$\nabla_\delta f(\alpha+\delta\mathbf{e}_n) = 1 - (\alpha^T + \delta\mathbf{e}_n^T)\mathbf{G}\mathbf{e}_n$$

$$\Rightarrow \delta_* = \frac{1-\alpha^T\mathbf{G}\mathbf{e}_n}{\mathbf{e}_n^T\mathbf{G}\mathbf{e}_n}$$

$$\Rightarrow \delta_* = \frac{1-y_n\mathbf{w}^T\mathbf{x}_n}{\mathbf{x}_n^T\mathbf{x}_n}$$

Since while doing co-ordinate ascent α_n^{new} should satisfy the constraint $0 \le \alpha_n^{new} \le C$,we would rewrite δ_* as:

$$\delta_* = \begin{cases} -\alpha_n^{old} & \alpha_n^{old} + \frac{1-y_n\mathbf{w}^T\mathbf{x}_n}{\mathbf{x}_n^T\mathbf{x}_n} < 0 \\ C - \alpha_n^{old} & \alpha_n^{old} + \frac{1-y_n\mathbf{w}^T\mathbf{x}_n}{\mathbf{x}_n^T\mathbf{x}_n} > C \\ \frac{1-y_n\mathbf{w}^T\mathbf{x}_n}{\mathbf{x}_n^T\mathbf{x}_n} & otherwise \end{cases}$$

Note that the new definition of δ satisfies both the constraint i.e. $0 \le \alpha_n^{new} \le C$ as well as $\arg\max_\delta f(\alpha+\delta\mathbf{e}_n)$ within this range of α as the objective function is concave and maxima lies either on corner point or where gradient is zero.

Co-ordinate ascent Algorithm:

1.Initialize all the $\alpha_n \forall$n ϵ(1,2,...N) in the range (0,C).

2.Pick an α_n uniform randomly.

update it as $\alpha_n^{new} = \alpha_n^{old} + \delta_*$

where δ_* is defined as above.

3.Go to step 2 if not converged.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 2

Since sum of pairwise inter-cluster and intra-cluster distance is constant,so minimizing the intra-cluster distance is same as maximizing inter-cluster distance.More formally

$$\sum_{n,m} \mathbb{1}[f_n = f_m] \left\|\mathbf{x}_n - \mathbf{x}_m\right\|^2 + \sum_{n,m} \mathbb{1}[f_n \neq f_m] \left\|\mathbf{x}_n - \mathbf{x}_m\right\|^2 = \text{constant}$$

So minimizing one part of L.H.S. leads to maximization of the second part in L.H.S. implicitly.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 3

$$\log p(\mathbf{x}_n^{obs}, \mathbf{x}_n^{miss}|\theta) = \log p(\mathbf{x}_n|\theta)$$
$$= -\frac{1}{2}(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mu) - \frac{D}{2}\log 2\pi - \frac{1}{2}\log|\mathbf{\Sigma}|$$

Here $\theta = (\mu, \mathbf{\Sigma})$

$$\sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta) = -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \mu)^T \mathbf{\Sigma}^{-1}(\mathbf{x}_n - \mu) - \frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\mathbf{\Sigma}|$$

$$= -\frac{1}{2}trace(\mathbf{\Sigma}^{-1}\mathbf{S}_\mu) - \frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\mathbf{\Sigma}|$$

$$\text{where } \mathbf{S}_\mu = \sum_{n=1}^{N}(\mathbf{x}_n - \mu)(\mathbf{x}_n - \mu)^T$$

$$\text{Expected CLL} = \mathbf{E}\left[\log p(\mathbf{x}_n^{obs}, \mathbf{x}_n^{miss}|\theta)\right] = \mathbf{E}\left[\sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)\right]$$

$$\mathbf{E}\left[\sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta)\right] = -\frac{1}{2}trace(\mathbf{\Sigma}^{-1}\mathbf{E}[\mathbf{S}_\mu]) - \frac{ND}{2}\log 2\pi - \frac{N}{2}\log|\mathbf{\Sigma}|$$

$$\mathbf{E}[\mathbf{S}_\mu] = \sum_{n=1}^{N}\mathbf{E}\left[\mathbf{x}_n\mathbf{x}_n^T - \mathbf{x}_n\mu^T - \mu\mathbf{x}_n^T + \mu\mu^T\right]$$
$$= \sum_{n=1}^{N}\left[\mathbf{E}[\mathbf{x}_n\mathbf{x}_n^T] - 2\mu\mathbf{E}(\mathbf{x}_n)^T + \mu\mu^T\right]$$

Also the Posterior $p(\mathbf{x}_n^{miss}|\mathbf{x}_n^{obs}, \theta) = \mathcal{N}(\alpha_n, \beta_n)$

where $\alpha_n = \mu^{miss} + \mathbf{\Sigma}_{miss,obs}\mathbf{\Sigma}_{obs,obs}^{-1}(\mathbf{x}_n^{obs} - \mu^{obs})$ and $\beta_n = \mathbf{\Sigma}_{miss,miss} - \mathbf{\Sigma}_{miss,obs}\mathbf{\Sigma}_{obs,obs}^{-1}\mathbf{\Sigma}_{obs,miss}$ Here $(\mu^{miss}, \mathbf{\Sigma}_{miss,miss}, \mathbf{\Sigma}_{miss,obs}, \mathbf{\Sigma}_{obs,obs}, \mathbf{\Sigma}_{obs,miss}, \mu^{obs})$ can all be directly derived using corresponding missing indices of $\mathbf{x}_n$ and given $\mathbf{\Sigma}, \mu$.For example for a given $\mathbf{x}_n = (\mathbf{x}_n^{obs}, \mathbf{x}_n^{miss}), \mu = (\mu^{obs}, \mu^{miss})$where μ^{obs}, μ^{miss} are obtained using the corresponding observed and missing indices of $\mathbf{x}_n$,Similarily we can write

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{obs,obs} & \mathbf{\Sigma}_{obs,miss} \\ \mathbf{\Sigma}_{miss,obs} & \mathbf{\Sigma}_{miss,miss} \end{bmatrix}$$

Hence,

$$\mathbf{E}[\mathbf{x}_n] = [\mathbf{x}_n^{obs}, \alpha_n]$$
$$\mathbf{E}[\mathbf{x}_n\mathbf{x}_n^T] = \mathbf{E}\begin{bmatrix} \mathbf{x}_n^{obs}\mathbf{x}_n^{obs^T} & \mathbf{x}_n^{obs}\mathbf{x}_n^{miss^T} \\ \mathbf{x}_n^{miss}\mathbf{x}_n^{obs^T} & \mathbf{x}_n^{miss}\mathbf{x}_n^{miss^T} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{x}_n^{obs}\mathbf{x}_n^{obs^T} & \mathbf{x}_n^{obs}\mathbf{E}[\mathbf{x}_n^{miss}]^T \\ \mathbf{E}[\mathbf{x}_n^{miss}]\mathbf{x}_n^{obs^T} & \mathbf{E}[\mathbf{x}_n^{miss}\mathbf{x}_n^{miss^T}] \end{bmatrix}$$

where $\mathbf{E}(\mathbf{x}_n)^{miss} = \alpha_n$ and $\mathbf{E}(\mathbf{x}_n)^{miss^T} = \alpha_n^T$ and $\mathbf{E}[\mathbf{x}_n^{miss}\mathbf{x}_n^{miss^T}] = \alpha_n\alpha_n^T + \beta_n$

The EM Algorithm:

1.Initialize θ as $\theta^{(0)}$,set t=1.

2.E step: Get the posterior $p(\mathbf{x}_n^{miss(t)}|\mathbf{x}_n^{obs}, \theta^{(t-1)})$ as shown above for all $n\epsilon(1, 2, ..., N)$.Calculate the expectation $\mathbf{E}\left[\mathbf{x}_n^{miss(t)}\right] = \alpha_n^{(t-1)}$ and $\mathbf{E}[\mathbf{x}_n^{miss(t)}\mathbf{x}_n^{miss(t)^T}] = \alpha_n{}^{(t-1)}\alpha_n{}^{(t-1)^T} + \beta_n^{(t-1)}$ which is then used to calculate $\mathbf{E}[\mathbf{x}_n\mathbf{x}_n^T]$ and $\mathbf{E}[\mathbf{x}_n]$ as shown above.Then use it to get the Expected CLL.

3.M step: Now maximize the expected complete data log-likelihood w.r.t. θ i.e.:

$\theta^{(t)} = \arg\max_\theta \mathbf{E}(\sum_{n=1}^{N} \log p(\mathbf{x}_n^{obs}, \mathbf{x}_n^{miss(t)}|\theta)) = \arg\max_\theta \mathbf{E}(\sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta))$

$$\mu^{(t)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{E}[\mathbf{x}_n]$$

$$\Sigma^{(t)} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{E}\left[\mathbf{x}_n\mathbf{x}_n^T\right] - \mu^{(t)}\mu^{(t)^T}$$

4.If not yet converged,set t=t+1 and go to Step.2.

Note: All the expectations calculated in this question are w.r.t. posterior i.e. $p(\mathbf{x}_n^{miss}|\mathbf{x}_n^{obs}, \theta)$.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 4

Given the N labeled examples $\{(\mathbf{x}_n, y_n)\}_{n=1}^N$ do MLE estimate to initialize θ as $\theta^{(0)}$ for EM.Here $\theta = \{\pi_k, \mu_k, \mathbf{\Sigma}_k\}_{k=1}^K$.So,

$$\pi_k^{(0)} = \frac{1}{N}\sum_{n=1}^{N} y_{nk}$$

$$\mu_k^{(0)} = \frac{1}{N_k}\sum_{n=1}^{N} y_{nk}\mathbf{x}_n$$

$$\mathbf{\Sigma}_k^{(0)} = \frac{1}{N_k}\sum_{n=1}^{N} y_{nk}(\mathbf{x}_n - \mu_k^{(0)})^T(\mathbf{x}_n - \mu_k^{(0)})$$

Here $N_k = \sum_{n=1}^{N} y_{nk}$ (used the expressions from slides).
Now in the given problem setting our latent variable will be each of the y_n for n ϵ $(N+1, ...N+M)$.
Expected Complete data log-likelihood:

$$\mathbf{E}\left[\log p(\mathbf{X}, \mathbf{y}|\theta)\right] = \sum_{n=N+1}^{N+M}\sum_{k=1}^{K}\gamma_{nk}\left[\log \pi_k + \log\left(\mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\right)\right] + \sum_{n=1}^{N} z_{nk}\left[\log \pi_k + \log\left(\mathcal{N}(\mathbf{x}_n|\mu_k, \Sigma_k)\right)\right]$$

The EM Algorithm:
1.Initialize θ as $\theta^{(0)}$ calculated above,set t=1.
2.E Step: Compute the expectation of each y_n given current parameters $\theta^{(t-1)}$.

$$\mathbf{E}[y_{nk}^{(t)}] = \gamma_{nk}^{(t)} = \frac{\pi_k^{(t-1)}\mathcal{N}(\mathbf{x}_n|\mu_k^{(t-1)}, \mathbf{\Sigma}_k^{(t-1)})}{\sum_{l=1}^{K}\pi_l^{(t-1)}\mathcal{N}(\mathbf{x}_n|\mu_l^{(t-1)}, \mathbf{\Sigma}_l^{(t-1)})} \quad \forall n\epsilon(N+1, ..., N+M) \text{ and } \forall k$$

3.M Step: Given responsibility $\gamma_{nk}^{(t)}$,we re-estimate θ via MLE for all k:

$$\mu_k^{(t)} = \frac{1}{N_k^{(t)}}\left\{\sum_{n=1}^{N} y_{nk}\mathbf{x}_n + \sum_{n=N+1}^{N+M}\gamma_{nk}^{(t)}\mathbf{x}_n\right\}$$

$$\mathbf{\Sigma}_k^{(t)} = \frac{1}{N_k^{(t)}}\left\{\sum_{n=1}^{N} y_{nk}(\mathbf{x}_n - \mu_k^{(t)})^T(\mathbf{x}_n - \mu_k^{(t)}) + \sum_{n=N+1}^{N+M}\gamma_{nk}^{(t)}(\mathbf{x}_n - \mu_k^{(t)})^T(\mathbf{x}_n - \mu_k^{(t)})\right\}$$

$$\pi_k^{(t)} = \frac{N_k^{(t)}}{N+M}$$

Here $N_k^{(t)} = \left\{\sum_{n=1}^{N} y_{nk} + \sum_{n=N+1}^{N+M}\gamma_{nk}^{(t)}\right\}$
4.Set t=t+1.Go to Step 2 if not converged yet.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 3

Student Name: Subham Kumar
Roll Number: 160707
Date: February 23, 2019

QUESTION 5

$$p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \theta) = \prod_{n=1}^{N} p(y_n, z_n|\mathbf{x}_n, \theta)$$

$$= \prod_{n=1}^{N} p(y_n|z_n, \mathbf{x}_n, \theta) p(z_n|\mathbf{x}_n, \theta)$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\mathcal{N}(\mathbf{w}_k^T \mathbf{x}_n, \beta^{-1})\pi_k\right)^{z_{nk}}$$

Here $\theta = \{(\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K), (\pi_1, \pi_2, ..., \pi_K)\}$. So Complete data log-likelihood can be written as:

$$\log p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \theta) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left[\log \pi_k - \frac{\beta}{2}(y_n - \mathbf{w}_k^T \mathbf{x}_n)^2 + constant\right]$$

Setting gradient of this CLL to zero for MLE estimate of θ w.r.t. π_k and $\mathbf{w}_k \forall k \epsilon (1, 2, ...K)$,

$$\hat{\pi}_k = \frac{N_k}{N} \text{ where } N_k = \sum_{n=1}^{N} z_{nk}$$

$$\hat{\mathbf{w}}_k = (\sum_{n=1}^{N} z_{nk} \mathbf{x}_n \mathbf{x}_n^T)^{-1} (\sum_{n=1}^{N} z_{nk} y_n \mathbf{x}_n)$$

The given model simply seems to be a mixture of K different least-square regression as the expression for $\hat{\mathbf{w}}_k$ same as that in least-square regression.We may want to use this model in case we have data such that they can be fit locally using linear regression models (overall can't be fit by just standard linear model due to non-linearity)i.e. if we have K line-clusters then each cluster can be modeled locally using this unregularized linear regression.

ALT-OPT Algorithm:

1.Initialize θ as $\theta^{(0)}$,set t=1.

2.$\forall n \epsilon (1, 2, ..., N)$

$$\hat{z}_n^{(t)} = \underset{k \epsilon (1,2,...,K)}{\arg\max}\ p(y_n, z_n = k|\mathbf{x}_n, \theta^{(t-1)})$$

$$= \underset{k \epsilon (1,2,...,K)}{\arg\max}\ p(z_n = k|\mathbf{x}_n, \theta^{(t-1)}) p(y_n|z_n = k, \mathbf{x}_n, \theta^{(t-1)})$$

$$= \underset{k \epsilon (1,2,...,K)}{\arg\max}\ \pi_k^{(t-1)} \mathcal{N}({\mathbf{w}_k^{(t-1)}}^T \mathbf{x}_n, \beta^{-1})$$

3.Given $\hat{\mathbf{Z}}^{(t)} = \{\hat{z}_1^{(t)}, \hat{z}_2^{(t)}, ..., \hat{z}_K^{(t)}\}$ re-estimate $\theta^{(t)}$ using MLE as:

$$\hat{\pi}_k^{(t)} = \frac{N_k^{(t)}}{N} \text{ where } N_k^{(t)} = \sum_{n=1}^{N} z_{nk}^{(t)}$$

$$\hat{\mathbf{w}}_k^{(t)} = \left(\sum_{n=1}^{N} z_{nk}^{(t)} \mathbf{x}_n \mathbf{x}_n^T\right)^{-1} \left(\sum_{n=1}^{N} z_{nk}^{(t)} y_n \mathbf{x}_n\right)$$

4.Go to step2 if not converged yet,set t=t+1.
If $\pi_k = \frac{1}{K}$ then expression for update of each z_n becomes:

$$\hat{z}_n = \underset{k\epsilon(1,2,...,K)}{\arg\max}\ \mathcal{N}(\mathbf{w}_k^T \mathbf{x}_n, \beta^{-1})$$

$$= \underset{k\epsilon(1,2,...,K)}{\arg\min}\ (y_n - \mathbf{w}_k^T \mathbf{x}_n)^2$$

So geometrically we have K hyperplanes $\mathbf{w}_k^T x_n$ for each $\mathbf{x}_n$ and $k\epsilon(1, 2, ..., K)$ and we are assigning z_n to the line-cluster for which the one point absolute error between y_n and $\mathbf{w}_k^T x_n$ is minimum.
Expected CLL:

$$\mathbf{E}\left[\log p(\mathbf{y}, \mathbf{z}|\mathbf{X}, \theta)\right] = \sum_{n=1}^{N}\sum_{k=1}^{K} \mathbf{E}[z_{nk}] \left[\log \pi_k - \frac{\beta}{2}(y_n - \mathbf{w}_k^T \mathbf{x}_n)^2 + constant\right]$$

The EM Algorithm:
1.Initialize θ as $\theta^{(0)}$,set t=1.
2.E step:

$$\mathbf{E}[z_{nk}^{(t)}] = \gamma_{nk}^{(t)} = \frac{\pi_k^{(t-1)} \exp\left(-\frac{\beta}{2}(y_n - \mathbf{w}_k^{(t-1)^T} \mathbf{x}_n)^2\right)}{\sum_{l=1}^{K} \pi_l^{(t-1)} \exp\left(-\frac{\beta}{2}(y_n - \mathbf{w}_l^{(t-1)^T} \mathbf{x}_n)^2\right)} \forall n, k$$

3.M step: Given responsibilities $\gamma_{nk}^{(t)}$ and $N_k^{(t)} = \sum_{n=1}^{N} \gamma_{nk}^{(t)}$ re-estimate θ via MLE

$$\hat{\pi}_k^{(t)} = \frac{N_k^{(t)}}{N}$$

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

4.Go to step 2 if not converged yet,set t=t+1.
Note that when $\beta \to \infty$,the value of the expression $\exp\left(-\frac{\beta}{2}(y_n - \mathbf{w}_l^{(t-1)^T} \mathbf{x}_n)^2\right)$ degrades slowly to zero in denominator for the l (call it l_0) which has smallest $\left|y_n - \mathbf{w}_l^{(t-1)^T} \mathbf{x}_n\right|$ as compared to others and hence $\gamma_{nl_0}^{(t)} \to 1$ as they cancel out to unity both in numerator and denominator as $\beta \to \infty$ since other terms goes to zero in denominator.This gives us a hard assignment irrespective of value of π_k.Similarily in ALT-OPT the ratio $\frac{\pi_k^{(t-1)} \mathcal{N}(\mathbf{w}_k^{(t-1)^T} \mathbf{x}_n, \beta^{-1})}{\pi_{l_0}^{(t-1)} \mathcal{N}(\mathbf{w}_{l_0}^{(t-1)^T} \mathbf{x}_n, \beta^{-1})}$ goes to zero $\forall k \neq l_0$ as $\beta \to \infty$ implying $\pi_{l_0}^{(t-1)} \mathcal{N}(\mathbf{w}_{l_0}^{(t-1)^T} \mathbf{x}_n, \beta^{-1})$is maximum and hence $z_{nl_0} = 1$.So in the limit $\beta \to \infty$ EM reduces to ALT-OPT.