Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: November 17, 2018

QUESTION 1

v is the eigen vector obtained from $\frac{1}{N}XX^T$
$XX^Tv = \lambda v$

Multiplying X^T on both sides,

$(X^TX)X^Tv = \lambda(X^Tv)$

Therefore, we can say that eigen vector u obtained from $\frac{1}{N}X^TX$ is X^T times of eigen vector v obtained from $\frac{1}{N}XX^T$.

We can derive the eigen vector u from the eigen vector v by X^Tv.

The advantage of computing eigen vector u from v is that v is calculated from $\frac{1}{N}XX^T$ which is a $N * N$ matrix. The eigen vector u if obtained directly should be obtained from $\frac{1}{N}X^TX$ which is $D * D$ matrix.Since, $D > N$ the computation time will be greater if we obtain eigen vector from $D * D$ matrix than computing eigen vector from $N * N$ vector which is smaller in size. The the advantage of this way of obtaining the eigen vectors of S is that it makes computing faster.

Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: November 17, 2018

QUESTION 2

The activation function $h(x) = x\sigma(\beta x)$

$h(x) = \frac{x}{1+exp(-\beta x)}$

If $\beta = 0$ then,

$h(x) = \frac{x}{1+exp(-0x)}$

$h(x) = \frac{x}{1+exp(0)}$

$h(x) = \frac{x}{1+1}$

$h(x) = \frac{x}{2}$

Therefore, for $\beta = 0$ the activation function $h(x)$ can approximate a linear function since $\frac{x}{2}$ is a linear function.

If $\beta = \infty$ and $x > 0$ then,

$h(x) = \frac{x}{1+exp(-\infty x)}$

$h(x) = \frac{x}{1+exp(-\infty)}$

$h(x) = \frac{x}{1+0}$

$h(x) = x$

If $\beta = \infty$ and $x < 0$ then,

$h(x) = \frac{x}{1+exp((-\infty)(-x))}$

$h(x) = \frac{x}{1+exp(\infty)}$

$h(x) = \frac{x}{1+\infty}$

$h(x) = 0$

Therefore, for $\beta = \infty$ the activation function $h(x)$ can approximate a relu function.

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: November 17, 2018

QUESTION 3

Given: z_n follows Multinoulli Distribution $(\pi_1, \pi_2, \ldots, \pi_k)$, $p(z_n) = \pi_k$

$p(y_n|x_n, z_n, w) = (\sigma(w_{zn}^T x_n))^{y_n}((1 - \sigma(w_{zn}^T x_n))^{1-y_n})$

$p(y_n = 1|x_n) = \sum_{k=1}^{K} p(y_n = 1|z_n = k, x_n, w)p(z_n)$

Since, y_n is Bernoulli,

$p(y_n = 1|x_n) = \sum_{k=1}^{K} \sigma(w_{zn}^T x_n)^{y_n} \pi_k$

$p(y_n = 1|x_n) = \sum_{k=1}^{K} \sigma(w_{zn}^T x_n) \pi_k$

In neural network,

The input layer is x_n

Activation layer is: $\sigma(w_{zn}^T x_n)$

Output Layer is: $\sum_{k=1}^{K} \sigma(w_{zn}^T x_n) \pi_k$

The connection parameters are: $(w_1, w_2, \ldots, w_k)$ and $(\pi_1, \pi_2, \ldots, \pi_k)$

Introduction to ML (CS771), Autumn 2018
Indian Institute of Technology Kanpur
Homework Assignment Number 4

Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: November 17, 2018

QUESTION 4

$p(X_{nm}|u_n, v_m, \theta_n, \phi_m) = N(X_{nm}|\theta_n + \phi_m + u_n^T v_m, \lambda_x^{-1})$

We need to compute the parameters, $\Theta = (u_n, v_m, \theta_n, \phi_m, W_u, W_v)$

$p(\Theta|X_{nm}) = p(X_{nm}|\Theta)p(\Theta)$

$p(\Theta|X_{nm}) = p(X_{nm}|\Theta)p(u_n)p(v_m)$

$p(\Theta|X_{nm}) = N(X_{nm}|\theta_n + \phi_m + u_n^T v_m, \lambda_x^{-1})N(u_n|W_u a_n, \lambda_u^{-1} I_k)N(v_m|W_v b_m, \lambda_v^{-1} I_k)$

The MAP objective is:

MAP=$log\Pi_{\Omega rn}\Pi_{\Omega_{cm}N(X_{nm}|\theta_n+\phi_m+u_n^T v_m^T,\lambda_x^{-1})N(u_n|W_u a_n,\lambda_u^{-1}I_k)N(v_m|W_v b_m,\lambda_v^{-1}I_k)}$

MAP=$\sum_{\Omega rn}\sum_{\Omega_{cm}log(N(X_{nm}|\theta_n+\phi_m+u_n^T v_m,\lambda_x^{-1})N(u_n|W_u a_n,\lambda_u^{-1}I_k)N(v_m|W_v b_m,\lambda_v^{-1}I_k)})$

MAP=$\sum_{\Omega rn}\sum_{\Omega_{cm}} \frac{-\lambda_x}{2}(X_{nm} - (\theta_n + \phi_m + u_n^T v_m))^2 - \frac{\lambda_u}{2}(u_n - W_u a_n)^2 - \frac{\lambda_v}{2}(v_m - W_v b_m)^2$

MAP=$\sum_{\Omega rn}\sum_{\Omega_{cm}} \frac{-\lambda_x}{2}(X_{nm} - (\theta_n + \phi_m + u_n^T v_m))^2 - \frac{\lambda_u}{2}(u_n - W_u a_n)^T(u_n - W_u a_n) - \frac{\lambda_v}{2}(v_m - W_v b_m)^T(v_m - W_v b_m)$

Differentiating MAP with respect to θ_n and equating to 0,

$\theta_n = \frac{\sum_{m\epsilon\Omega_{rn}} X_{nm} - (\phi_m + u_n^T v_m)}{|\Omega_{rn}|}$

Differentiating MAP with respect to ϕ_m and equating to 0,

$\phi_m = \frac{\sum_{n\epsilon\Omega_{cm}} X_{nm} - (\theta_n + u_n^T v_m)}{|\Omega_{cm}|}$

Differentiating MAP with respect to u_n and equating to 0,

$\sum_{m\epsilon\Omega_{rn}} \lambda_x v_m(X_{nm} - (\theta_n + \phi_m + u_n^T v_m)) - \lambda_u(u_n - W_u a_n) = 0$

$\sum_{m\epsilon\Omega_{rn}} \lambda_x v_m X_{nm} - \lambda_x v_m \theta_n - \lambda_x v_m \phi_m - \lambda_x v_m v_m^T u_n - \lambda_u u_n + \lambda_u W_u a_n = 0$

$\sum_{m\epsilon\Omega_{rn}} \lambda_x v_m X_{nm} - \lambda_x v_m \theta_n - \lambda_x v_m \phi_m + \lambda_u W_u a_n = \sum_{m\epsilon\Omega_{rn}} \lambda_x v_m v_m^T u_n + \lambda_u u_n$

$u_n = \sum_{m\epsilon\Omega_{rn}}(\lambda_x v_m v_m^T + \lambda_u I_k)^{-1}(\lambda_x v_m X_{nm} - \lambda_x v_m \theta_n - \lambda_x v_m \phi_m + \lambda_u W_u a_n)$

Differentiating MAP with respect to v_m and equating to 0,

$v_m = \sum_{n \epsilon \Omega_{cm}} (\lambda_x u_n u_n^T + \lambda_v I_k)^{-1} (\lambda_x u_n X_{nm} - \lambda_x u_n \theta_n - \lambda_x u_n \phi_m + \lambda_v W_v b_m)$

Differentiating MAP with respect to W_u and equating to 0,

$\frac{\lambda_u}{2}(-2u_n a_n^T + W_u(2a_n a_n^T)) = 0$

$W_u = \sum_{n=1}^{N} (u_n a_n^T))(a_n a_n^T)^{-1}$

Differentiating MAP with respect to W_v and equating to 0,

$W_v = \sum_{m=1}^{M} (v_m b_m^T))(b_m b_m^T)^{-1}$

ALT-OPT Algo:

Step 1: Initialize Θ as $\Theta^{(0)}$

Step 2: Update each element of Θ with the appropriate equations.

Step3: Repeat till converge.
The loss function is:

$-\sum_{\Omega rn} \sum_{\Omega_{cm}} \frac{-\lambda_x}{2}(X_{nm} - (\theta_n + \phi_m + u_n^T v_m))^2 - \frac{\lambda_u}{2}(u_n - W_u a_n)^T (u_n - W_u a_n) - \frac{\lambda_v}{2}(v_m - W_v b_m)^T (v_m - W_v b_m)$

Student Name: Aneet Kumar Dutta
Roll Number: 18111401
Date: November 17, 2018

QUESTION 5

We observed that the reconstruction gets better with increase in the number of K because the number of features used to reconstruct is increased and hence information retained is more.

For plots please click the link:

Please click here:

Programming part2:

Visually t-SNE works better for clustering task.

For plots please click the link:

Please click here