**CS 6140 Machine Learning**
Northeastern University, Fall 2018

# Homework 4

Name: Ruchitha Midigarahalli Shanmugha Sundar, UID: 001838207

## 1 Problem 1

### 1.1

If A and B are n*n square matrices then

$$Tr(AB) = \sum_{i=1}^{N}(AB)_{ii}$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{N} A_{ij}B_{ji}$$
$$= \sum_{j=1}^{N}\sum_{i=1}^{N} B_{ji}A_{ij}$$
$$= \sum_{j=1}^{N}(BA)_{ij}$$
$$= Tr(BA)$$

We can extend this for A, B and C

$$Tr(ABC) = \sum_{i} (ABC)_{ii}$$

Using ABC = A(BC) we have:

$$= \sum_{i}\sum_{j} A_{ij}(BC)_{ji} = \sum_{i}\sum_{j}\sum_{k} A_{ij}B_{jk}C_{ki}$$

Now we can move the matrices in cyclic order

$$= \sum_{k}\sum_{i}\sum_{j} C_{ki}A_{ij}B_{jk} = Tr(CAB)$$
$$= \sum_{j}\sum_{k}\sum_{i} B_{jk}C_{ki}A_{ij} = Tr(BCA)$$

Therefore, Tr(ABC) = Tr(BCA) = Tr(CAB).
This can be extended to any number of square matrices. Hence Trace is invariant under cyclic permutation.

1.2
Solution-1.2.1:
The first component vector(v) :
Select a vector whose values sum to 1 when squared

$$V = [1/\sqrt{2}, 1/\sqrt{2}]^T$$

Solution-1.2.2: The co-ordinates in 1-D space obtained after projecting points into 1-D space using the first component vector is as follows:
Let z be the point in 1-D space, x be the point in 2-D space. we have :

$$z = x^T * v$$

$$\text{First Point } x_1^T = [-1, -1], z_1 = [-1, -1][1/\sqrt{2}, 1/\sqrt{2}]^T = -\sqrt{2}$$

$$\text{Second Point } x_2^T = [0, 0], z_2 = [0, 0][1/\sqrt{2}, 1/\sqrt{2}]^T = 0$$

$$\text{Third Point } x_3^T = [1, 1], z_3 = [1, 1][1/\sqrt{2}, 1/\sqrt{2}]^T = \sqrt{2}$$

Solution-1.2.3:
Mean of the projected data, $\mu = (-\sqrt{2} + 0 + \sqrt{2})/3 = 0$
Variance of the Projected data

$$\sigma^2 = \frac{\sum(z_i - \mu)^2}{N} = \frac{1}{3}((-\sqrt{2} - 0)^2 + (0 - 0)^2 + (\sqrt{2} - 0)^2) = \frac{4}{3}$$

Solution-1.2.4:
Equation to reconstruct original points: x = z * v

$$\text{First point}, x_1 = z_1 * v = -\sqrt{2} * [1/\sqrt{2}, 1/\sqrt{2}] = [-1, -1]$$

$$\text{Second point}, x_2 = z_1 * v = 0 * [1/\sqrt{2}, 1/\sqrt{2}] = [0, 0]$$

$$\text{Third point}, x_3 = z_1 * v = -\sqrt{2} * [1/\sqrt{2}, 1/\sqrt{2}] = [1, 1]$$

we can see from the above values , Reconstruction Error = 0

# 2 Problem 2

MATRIX FACTORIZATION:
Solution-2.1: Regularized Squared Error is as given below:

$$argmin_{u,v} \frac{\lambda}{2}(||U||_F^2 + ||V||_F^2) + \frac{1}{2}\sum_{i,j}(y_{ij} - u_i^T v_j)^2$$

Gradient of the above regularized squared error w.r.t $u_i$ is as follows:
Taking derivative w.r.t $u_i$ we have :

$$\delta_{u_i} = \lambda U + \sum_{i,j}(y_{ij} - u_i^T v_j) * (-v_j)$$

$$\delta_{v_j} = \lambda V + \sum_{i,j} (y_{ij} - u_i^T v_j) * (-u_i)$$

Solution-2.2:
In Alternate least square, we first fix V and solve for optimal value of U by setting $\delta_{v_i}$ to zero:

$$\delta_{u_i} = \lambda U + \sum_{i,j} (y_{ij} - u_i^T v_j) * (-v_j) = 0$$

Solving the above equation for $u_i$

$$u_i = \left( \lambda I_k + \sum_j v_j v_j^T \right)^{-1} \sum_j y_{ij} v_j$$

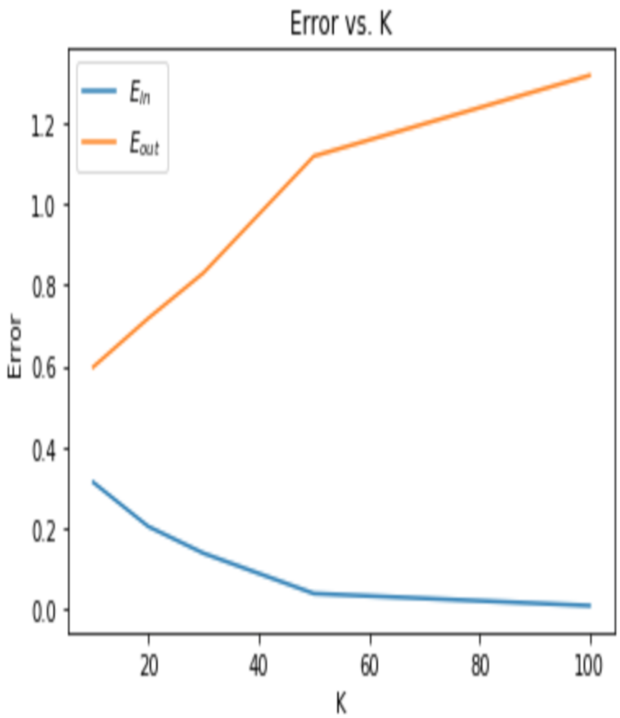
Now, we fix U and solve for V

$$\delta_{v_j} = \lambda V + \sum_{i,j} (y_{ij} - u_i^T v_j) * (-u_i) = 0$$

Solving the above equation for $v_j$

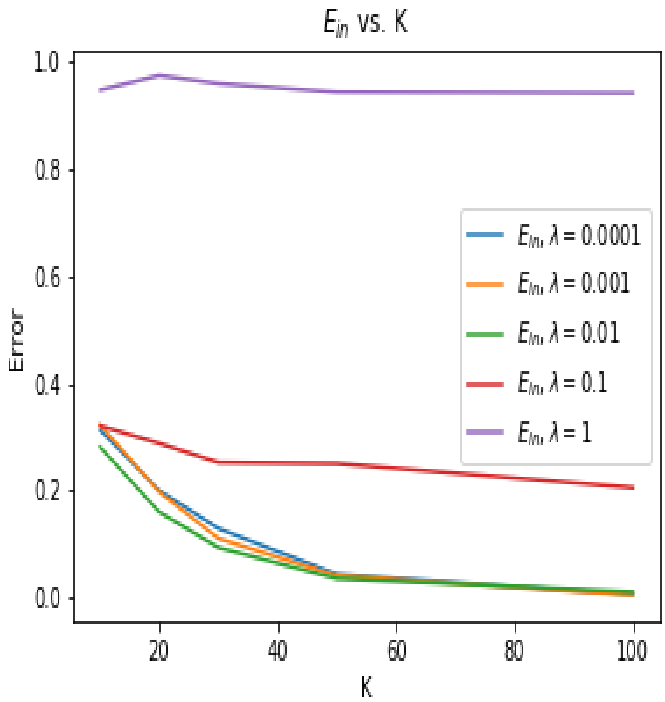
$$v_j = \left( \lambda I_k + \sum_i u_i u_i^T \right)^{-1} \sum_i y_{ij} u_i$$

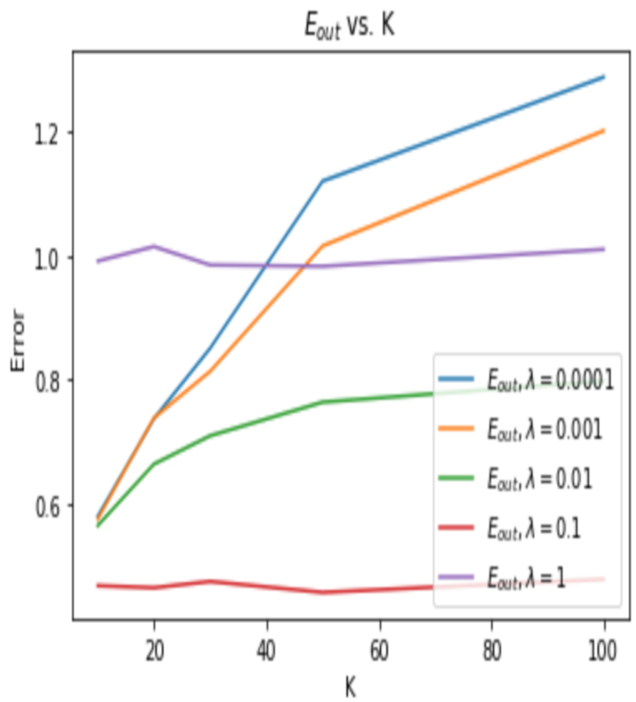
Solution-2.3:
python code and Jupyter notebook both submitted
Solution-2.4:
python code submitted



From the graph above we can notice that with the increase in K, the testing error increases due to over-fitting whereas with increase in K, training error decreases because of reconstruction.

Solution-2.5:
python code submitted




We can see that the both training error and testing error is decreasing for most of the regularization term value( $\lambda$ ) but when the regularization is high( example when $\lambda$ =1) the error has increased as it reduces learning rate resulting in under-fitting model.

# 3 Problem 3

**Expectation Maximization:**
**Solution-3.1:**
Given K Bernoulli distributions with parameter vector $p^{(k)} \in (0,1)^D$
Distribution, $\pi = [\pi_0, \pi_1, \pi_2, ..., \pi_k]$
Let $p = [p^{(1)}, p^{(2)}, ...., p^{(k)}]$
Let $A_k$ be the event that occurs when $x = x^{(i)}$ is taken from $p^{(k)}$

$$\text{Probability of the data point, } P(x|p,\pi) = \sum_k P(x|A_k, p, \pi)P(A_k|p,\pi)$$
$$= \sum_k \pi_k P(x|p^{(k)})$$

**Solution-3.2:**

$$P(X, Z|\pi, p) = \Pi_{i=1}^N P(x^{(i)}, z^{(i)}|\pi, p)$$
$$= \Pi_{i=1}^N P(x^{(i)}|z^{(i)}, \pi, p)P(z^{(i)}|\pi)$$
$$= \Pi_{i=1}^N \left[\Pi_{k=1}^K \left[P(x^{(i)}|p^{(k)})\right]^{z_{k(i)}}\right]\left[\Pi_{k=1}^K \pi_k^{z_k^{(i)}}\right]$$

Now taking log on both sides we have:

$$logP(X,Z|p,\pi) = \sum_{i=1}^N \left[\sum_{k=1}^K z_{k(i)} \log\left[P(x^{(i)}|p^{(k)})\right]\right] + \left[\sum_{k=1}^K z_k^{(i)} \log \pi_k\right]$$
$$= \sum_{i=1}^N \sum_{k=1}^k z_k^{(i)} \log\left[\log P(x^{(i)}|p^{(k)}) + \log \pi_k\right]$$

Let $p \in (0,1)^D$ be the Bernoulli parameter resulting vector
Now using $P(x|p) = \Pi_{d=1}^D p_d^{x_d}(1-p_d)^{(1-x_d)}$ where $P(x_d = 1) = p_d$

$$= \sum_{i=1}^N \sum_{k=1}^K Z_k^{(i)} \left[\log \pi_k + \log \Pi_{d=1}^D (p_d^{(k)})^{x_d^{(i)}}(1-p_d^{(k)})^{1-x_d^{(i)}}\right]$$
$$\sum_{i=1}^N \sum_{k=1}^K Z_k^{(i)} \left[\log \pi_k + \sum_{d=1}^D \left[(x_d^{(i)} log(p_d^{(k)}) + (1-x_d^{(i)})\log(1-p_d^{(k)})\right]\right]$$

For $E[logP(X,Z|p,\pi)]$ substituting $E[z_k^{(i)}] = \eta(z_k^{(i)})$ we have

$$E[logP(X,Z|p,\pi)] = \sum_{i=1}^N \sum_{k=1}^K \eta(z_k^{(i)}) \left[\log \pi_k + \sum_{d=1}^D \left[(x_d^{(i)} log(p_d^{(k)}) + (1-x_d^{(i)})\log(1-p_d^{(k)})\right]\right] \text{ -> Equation 1}$$

**Solution-3.3:**
In order to get $p_d$ take derivative $E[logP(X,Z|p,\pi)]$ w.r.t $p_d$ and set it to zero:

$$\frac{\delta}{\delta p_d^{(k)}} E[logP(X,Z|p,\pi)] = \sum_{i=1}^{N} \eta(z_k^{(i)}) \left[ \frac{x_d^{(i)}}{p_d^{(k)}} + \frac{1-x_d^{(i)}}{1-p_d^{(k)}} \right] = 0$$

$$\sum_{i=1}^{N} \eta(z_k^{(i)}) \left[ x_d^{(i)}(1-p_d^{(k)}) + (1-x_d^{(i)})p_d^{(k)} \right] = 0$$

$$\sum_{i=1}^{N} \eta(z_k^{(i)}) \left[ x_d^{(i)} - p_d^{(k)} \right] = 0$$

Now solving for $p_d^{(k)}$ we have:

$$\begin{aligned} p_d^{(k)} &= \frac{\sum_{i=1}^{N} \eta(z_k^{(i)}) x_d^{(i)}}{\sum_{i=1}^{N} \eta(z_k^{(i)})} \\ &= \frac{\sum_{i=1}^{N} \eta(z_k^{(i)}) x_d^{(i)}}{N_k} \end{aligned}$$

In order to solve for $\pi_k$ we need to minimize just the first term of Equation 1 which is a function of $\pi$

$$L(\pi,\lambda) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \eta(z_k^{(i)}) log\pi_k + \lambda\Big(\sum_{k=1}^{K} \pi_k - 1\Big)$$

By taking derivative of $L(\pi,\lambda)$ w.r.t $\pi_k$

$$\frac{\delta}{\delta \pi_k} L(\pi,\lambda) = -\sum_{i=1}^{N} \frac{\eta(z_k^{(i)})}{\pi_k} + \lambda = 0$$

$$\pi_k = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)})}{N} + \lambda = 0$$

$$\pi_k = \frac{\sum_{i=1}^{N} \eta(z_k^{(i)})}{\lambda} = \frac{N_k}{\lambda} \text{ —-> equation 2}$$

Solving for $\lambda$:

$$L(\lambda) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \eta(z_k^{(i)})(logN_k - log\lambda) + \Big(\sum_{k=1}^{K} N_k - \lambda\Big)$$

on taking derivative w.r.t $\lambda$ we have:

$$\frac{1}{\lambda}\sum_{i=1}^{N}\sum_{k=1}^{K} \eta(z_k^{(i)}) - 1 = 0$$

$$\lambda = \sum_{i=1}^{N}\sum_{k=1}^{K} \eta(z_k^{(i)})$$

$$=> \lambda = \sum_{k=1}^{K} N_k \text{ —-> equation 3}$$

substituting 3 in 2 we have:

$$\pi_k = \frac{\sum_{i=1} N\eta(z_k^{(i)})}{\lambda} = \frac{N_k}{\sum_{k=1}^{K} N_k}$$

Hence Proved!