CS 7641 CSE/ISYE 6740 Homework 3

Prakash, Fall 2021

Deadline: 11/09 Tuesday, 11:59pm ET

- Submit your answers as an electronic copy on gradescope.
- No unapproved extension of deadline is allowed. Zero credit will be assigned for late submissions. Email request for late submission may not be replied.
- For typed answers with LaTeX (recommended) or word processors, extra credits will be given (= 5 points). If you handwrite, try to be clear as much as possible. No credit may be given to unreadable handwriting.
- Explicitly mention your collaborators if any.
- Recommended reading: PRML[1] Section 3.1, 3.2

1 Linear Regression [30 pts]

In class, we derived a closed form solution (normal equation) for linear regression problem: $\hat{\theta} = (X^T X)^{-1} X^T Y$. A probabilistic interpretation of linear regression tells us that we are relying on an assumption that each data point is actually sampled from a linear hyperplane, with some noise. The noise follows a zero-mean Gaussian distribution with constant variance. Specifically,

$$Y^i = \theta^T X^i + \epsilon^i \tag{1}$$

where $\epsilon^i \sim \mathcal{N}(0, \sigma^2 I)$, $\theta \in \mathbb{R}^d$, and $\{X^i, Y^i\}$ is the i-th data point. In other words, we are assuming that each every point is independent to each other and that every data point has same variance.

(a) Using the normal equation, and the model (Eqn. 1), derive the expectation $\mathbb{E}[\hat{\theta}]$. Note that here X is fixed, and only Y is random, i.e. "fixed design" as in statistics. [6 pts]

↪ We know from Eq.1 and the normal equation :

↪ $E[\hat{\theta}] = E[(X^T X)^{-1} X^T Y] = E[(X^T X)^{-1} X^T (X\theta + \epsilon)] = (X^T X)^{-1} X^T E[X\theta + \epsilon]$

↪ But, $E[A + BX] = A + BE[X]$, therefore, $E[\hat{\theta}] = E[(X^T X)^{-1} X^T X\theta] + X^{-1} X^{-T} X^T E[\epsilon]$

↪ $E[\hat{\theta}] = E[I\theta] + E[\epsilon] = E[\theta] + X^{-1} \cdot 0 = E[\theta] = \theta$

[1] Christopher M. Bishop, Pattern Recognition and Machine Learning, 2006, Springer.

(b) Similarly, derive the variance Var$[\hat{\theta}]$. [6 pts]

↪ $Var[\hat{\theta}] = Var[(X^TX)^{-1}X^TY]$

↪ We know,$Var[AB] = AVar[b]A^T$

↪ Therefore, $Var[\hat{\theta}] = ((X^TX)^{-1}X^T)Var(Y)((X^TX)^{-1}X^T)^T$

↪ $Var[\hat{\theta}] = ((X^TX)^{-1}X^T)Var(X\theta + \epsilon)((X^TX)^{-1}X^T)^T$

↪ $Var[\hat{\theta}] = ((X^TX)^{-1}X^T)(XVar(\theta) + Var(\epsilon))((X^TX)^{-1}X^T)^T$

↪ $Var[\hat{\theta}] = ((X^TX)^{-1}X^TXVar(\theta) + \sigma^2((X^TX)^{-1}X^T)((X^TX)^{-1}X^T)^T$

↪ Therefore, $Var[\hat{\theta}] = Var(\theta) + \sigma^2(X^TX)^{-1} = \sigma^2(X^TX)^{-1}$ Since θ is constant.

(c) Under the white noise assumption above, someone claims that $\hat{\theta}$ follows Gaussian distribution with mean and variance in (a) and (b), respectively. Do you agree with this claim? Why or why not? [8 pts]

↪ We know that $\hat{\theta} = (X^TX)^{-1}X^T(X\theta + \epsilon)$

↪ From this, $\hat{\theta} = A\theta + B\epsilon$ which is a linear combination of ϵ and θ, and only ϵ varies with a distribution (which is Gaussian).

↪ So a linear combination of a Gaussian is also a Gaussian, and from the previous steps, the mean and variance of Gaussian is θ and variance is $\sigma^2(X^TX)^{-1}$.

(d) Weighted linear regression

Suppose we keep the independence assumption but remove the same variance assumption. In other words, data points would be still sampled independently, but now they may have different variance σ_i. Thus, the covariance matrix of Y would be still diagonal, but with different values:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}. \tag{2}$$

Derive the estimator $\hat{\theta}$ (similar to the normal equations) for this problem using matrix-vector notations with Σ. [10 pts]

↪ We know that $p(\epsilon_i = y_i - \theta^T x_i \mid x_i; \theta) = \frac{1}{(2\pi)^{0.5}\sigma_i} exp \frac{-(y_i - \theta^T x_i)^2}{2\sigma_i^2}$

↪ Likelihood $L = \prod_{i=1}^{n} p(\epsilon_i \mid x_i; \theta)$

↪ Log Likelihood $LL = \frac{1}{2}\sum_{i=1}^{n} \log \frac{1}{2\pi\sigma_i^2} - \sum_{i=1}^{n} \frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}$

↪ We set $\frac{\partial LL}{\partial \theta} = 0$, we know that the first term is independent of θ and so our objective function is only dependent on $-\sum_{i=1}^{n} \frac{(y_i - \theta^T x_i)^2}{2\sigma_i^2}$ which can be rewritten in vector notation as $-(Y - X\theta)^T\Sigma^{-1}(Y - X\theta)$

↪ Therefore, $\frac{\partial LL}{\partial \theta} = \frac{\partial}{\partial \theta}(-(Y - X\theta)^T \Sigma^{-1}(Y - X\theta)) = 0$

↪ $\frac{\partial LL}{\partial \theta} = 2X^T \Sigma^{-1}(Y - X\theta) = 0$

↪ Hence, $X^T \Sigma^{-1}(Y - X\theta) = 0 \implies \mathrm{X}^T \Sigma^{-1} Y = X^T \Sigma^{-1} X\theta$

↪ Equivalently, $\hat{\theta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} Y$

2 Ridge Regression [15 pts]

For linear regression, it is often assumed that $y = \theta^\top \mathbf{x} + \epsilon$ where $\theta, \mathbf{x} \in \mathbb{R}^m$ by absorbing the constant term, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian random variable. Given n i.i.d samples $(\mathbf{x}^1, y^1), ..., (\mathbf{x}^n, y^n)$, we define $\mathbf{y} = (y^1, ..., y^n)^\top$ and $X = (\mathbf{x}^1, ..., \mathbf{x}^n)^\top$. Thus, we have $\mathbf{y} \sim \mathcal{N}(X\theta, \sigma^2 I)$. Show that the ridge regression estimate is the mean of the posterior distribution under a Gaussian prior $\theta \sim \mathcal{N}(0, \tau^2 I)$. Find the explicit relation between the regularization parameter λ in the ridge regression estimate of the parameter θ, and the variances σ^2, τ^2.

↪ It is given that $p(\theta) = \mathcal{N}(\theta \mid 0, I\tau^2)$ (Also called prior), $p(y \mid \theta) = \mathcal{N}(y \mid X\theta, I\sigma^2)$ (Likelihood), $p(y) = k$ (constant, normalizing factor).

↪ From Bayes' rule, posterior $p(\theta \mid y) = \frac{p(y \mid \theta)p(\theta)}{p(y)} = \frac{1}{k}\mathcal{N}(\theta \mid 0, \tau^2)\mathcal{N}(y \mid X\theta, I\sigma^2)$

↪ $p(\theta \mid y) = \frac{1}{k} \frac{1}{(2\pi\sigma^2)^{0.5n}} exp(-\frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2}) \frac{1}{(2\pi\sigma^2)^{0.5m}} exp(-\frac{\theta^T \theta}{2\tau^2})$

↪ Log likelihood $LL = \log p(\theta \mid y) = -\frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} - \frac{\theta^T \theta}{2\tau^2} + c$ (c is the constant part).

↪ Now, $\frac{\partial LL}{\partial \theta} = \frac{2X^T (y - X\theta)}{2\sigma^2} - \frac{2\theta}{2\tau^2} = 0$

↪ Therefore, $\hat{\theta} = (X^T X + \frac{\sigma^2}{\tau^2} I)^{-1} X^T y$.

- Comparing this to ridge regression estimate - $(X^T X + \lambda I)^{-1} X^T y$, we can set $\lambda = \frac{\sigma^2}{\tau^2}$
- Also, we know that the posterior also follows a gaussian (say $\mathcal{N}(\mu, \Sigma)$). We can say this as the prior and the likelihood both follow gaussian distributions and the product of gaussians is also a gaussian. We perform the upcoming analyses by comparing exponents.

↪ From the posterior, we have the exponent as $-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu)$. From the prior*likelihood, we have the exponent as $-\frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} - \frac{\theta^T \theta}{2\tau^2}$.

↪ We equate them to find μ, Σ.

↪ Therefore, $-\frac{1}{2}(\theta - \mu)^T \Sigma^{-1}(\theta - \mu) = -\frac{(y - X\theta)^T (y - X\theta)}{2\sigma^2} - \frac{\theta^T \theta}{2\tau^2}$

↪ Equivalently, $-\frac{1}{2}(\theta^T\Sigma^{-1}\theta - \theta^T\Sigma^{-1}\mu - \mu^T\Sigma^{-1}\theta + \mu^T\Sigma^{-1}\mu) = -\frac{(y-X\theta)^T(y-X\theta)}{2\sigma^2} - \frac{\theta^T\theta}{2\tau^2}$

- Comparing the coefficients of the second order of θ:

↪ $\theta^T\Sigma^{-1}\theta = \frac{1}{\sigma^2}\theta^T X^T X\theta + \frac{1}{\tau^2}\theta^T\theta = \frac{1}{\sigma^2}\theta^T(X^T X + \frac{\sigma^2}{\tau^2}I)\theta$

↪ Therefore, $\Sigma^{-1} = \frac{1}{\sigma^2}(X^T X + \frac{\sigma^2}{\tau^2}I)$

- Comparing the coefficients of the first order of θ:

↪ $\theta^T\Sigma^{-1}\mu = \frac{1}{\sigma^2}\theta^T X^T y$

↪ Therefore, $\mu = (X^T X + \frac{\sigma^2}{\tau^2}I)X^T y$

- We clearly see that the mean is equivalent to the ridge regression estimate with the regularization parameter λ being equal to $\frac{\sigma^2}{\tau^2}$.

3 Bayes Classifier

3.1 Bayes Classifier With General Loss Function

In class, we talked about the popular 0-1 loss function in which $L(a,b) = 1$ for $a \neq b$ and 0 otherwise, which means all wrong predictions cause equal loss. Yet, in many other cases including cancer detection, the asymmetric loss is often preferred (misdiagnosing cancer as no-cancer is much worse). In this problem, we assume to have such an asymmetric loss function where $L(a,a) = L(b,b) = 0$ and $L(a,b) = p, L(b,a) = q, p \neq q$. Write down the the Bayes classifier $f : X \to Y$ for binary classification $Y \in \{-1, +1\}$. Simplify the classification rule as much as you can. [20 pts]

- We obtain the following for $L(Y_1, Y_2)$:

$$\begin{cases} 0 & Y_1 = Y_2 (= 1 \text{ or } -1) \\ p & Y_1 = 1, Y_2 = -1 \\ q & Y_2 = 1, Y_1 = -1 \end{cases}$$

↪ Posterior probability of a test point is $p(Y = i \mid X) = \frac{p(X \mid Y)p(Y)}{p(X)}$

- From the above, we observe as follows:

↪ For $i = 1$: $L_1 = p(Y_1 = 1 \mid X)L(Y_1 = 1, Y_2 = 1) + p(Y_1 = 1 \mid X)L(Y_1 = 1, Y_2 = -1) = p(Y_1 = 1 \mid X)p$

↪ For $i = -1$: $L_{-1} = p(Y_1 = -1 \mid X)L(Y_1 = -1, Y_2 = -1) + p(Y_1 = -1 \mid X)L(Y_1 = -1, Y_2 = 1) = p(Y_1 = -1 \mid X)q$

- To find a bayes classifier rule for $f : X \to Y$ for $Y \in \{-1, +1\}$ with the bayes decision rule as follows :

↪ $\frac{L_1(X)}{L_{-1}(X)} = \frac{p(Y_1 = 1 \mid X)p}{p(Y_1 = -1 \mid X)q} = \frac{p(Y_1 = 1)p(X \mid Y_1 = 1)p}{p(Y_1 = -1)p(X \mid Y_1 = -1)q}$

↪ We also know that the testing point belongs to '+1' classification if $\frac{L_1(X)}{L_{-1}(X)} < 1$ and '-1' classification if $\frac{L_1(X)}{L_{-1}(X)} > 1$.

- Thus $f(X) =$

$$\begin{cases} 1 & p(Y_1 = 1 \mid X)p > p(Y_1 = -1 \mid X)q \\ -1 & p(Y_1 = 1 \mid X)p < p(Y_1 = -1 \mid X)q \end{cases}$$

3.2 Gaussian Class Conditional distribution

(a) Suppose the class conditional distribution is a Gaussian. Based on the general loss function in problem 3.1, write the Bayes classifier as $f(X) = \text{sign}(h(X))$ and simplify h as much as possible. What is the geometric shape of the decision boundary? [10 pts]

$\hookrightarrow$ From 3.1 we define $g(X) = \dfrac{p(Y_1 = 1)p(X \mid Y_1 = 1)p}{p(Y_1 = -1)p(X \mid Y_1 = -1)q}$

$\hookrightarrow$ $g(x) = \dfrac{p(Y_1 = 1)p(X \mid \mu_1, \Sigma_1)p}{p(Y_1 = -1)p(X \mid \mu_{-1}, \Sigma_{-1})q}$

$\hookrightarrow$ $g(x) = \dfrac{p(Y_1 = 1)\Sigma_{-1}^{-0.5}}{p(Y_1 = -1)\Sigma_1^{-.5}} \exp(\dfrac{1}{2} - (X - \mu_1)^T\Sigma_1^{-1}(X - \mu_1) + (X - \mu_{-1})^T\Sigma_{-1}^{-1}(X - \mu_{-1}))$

$\hookrightarrow$ If $h(X) > 0$, then $f(X) = sign(h(X)) = 1$. Else if $h(X) < 0$ then $f(X) = sign(h(X)) = -1$. Thus, with $h(X) = \log X$ we have : $g(X) > 1 \implies h(X) > 0, f(X) = 1$ and $g(X) < 1 \implies h(X) < 0, f(X) = -1$.

- The geometric decision boundary is as follows:

$\hookrightarrow$ $h(X) = \log(g(X)) = \log \dfrac{p(Y_1 = 1)\Sigma_{-1}^{-0.5}}{p(Y_1 = -1)\Sigma_1^{-.5}} - \dfrac{1}{2} - (X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1) + (X-\mu_{-1})^T\Sigma_{-1}^{-1}(X-\mu_{-1}) = 0$

$\hookrightarrow$ Clearly, if $\Sigma_1 \neq \Sigma_{-1}$, the boundary will be quadratic in terms of X. Thus, the shape of the geometric decision boundary is an n-dimensional quadratic surface.

(b) Repeat (a) but assume the two Gaussians have identical covariance matrices. What is the geometric shape of the decision boundary? [10 pts]

- If the two covariance matrices of both the Gaussians are identical, then the second order coefficients of X are eliminated. ($\Sigma_1 = \Sigma_{-1}$)

$\hookrightarrow$ $h(X) = c + \dfrac{1}{2}[((\mu_1^T - \mu_{-1}^T)\Sigma_1^{-1}X) + X^T(\mu_1 - \mu_{-1})\Sigma_1^{-1}] + \mu_1^T\Sigma_1^{-1}\mu_1 - \mu_{-1}^T\Sigma_1^{-1}\mu_{-1}$

$\hookrightarrow$ Hence the shape of the decision boundary is an n-dimensional plane.

(c) Repeat (a) but assume now that the two Gaussians have covariance matrix which is equal to the identity matrix. What is the geometric shape of the decision boundary? [10 pts]

- If the two covariance matrices of both the Gaussians are identical (with $\Sigma_1 = \Sigma_{-1} = I$) :

$\hookrightarrow$ $h(X) = c + \dfrac{1}{2}[(\mu_1^T - \mu_{-1}^T)X + X^T(\mu_1 - \mu_{-1})] = c + (\mu_1^T - \mu_{-1}^T)X$

$\hookrightarrow$ Thus, the shape of the decision boundary is an n-dimensional plane that is orthogonal to $(\mu_1^T - \mu_{-1}^T)$.

4 Logistic Regression

Logistic regression is named after the log-odds of success (the logit of the probability) defined as below:

$$\ln\left(\frac{P[Y=1|X=x]}{P[Y=0|X=x]}\right)$$

where

$$P[Y=1|X=x] = \frac{\exp(w_0 + w^T x)}{1+\exp(w_0 + w^T x)}$$

(a) Show that log-odds of success is a linear function of X. [6 pts]

↪ We have

$$P[Y=1|X=x] = \frac{\exp(w_0 + w^T x)}{1+\exp(w_0 + w^T x)}$$

↪ Also,

$$P[Y=0|X=x] = 1 - \frac{\exp(w_0 + w^T x)}{1+\exp(w_0 + w^T x)} = \frac{1}{1+\exp(w_0 + w^T x)}$$

↪ Therefore, $\ln\left(\frac{P[Y=1|X=x]}{P[Y=0|X=x]}\right) = ln\left(\exp(w_0 + w^T x)\right) = w_0 + w^T x$

↪ Clearly, the above is a linear function of x.

(b) Show that the logistic loss $L(z) = \log\left(1+\exp(-z)\right)$ is a convex function. [9 pts]

↪ The derivative $L^{'}(z) = \frac{\partial L}{\partial z} = -\frac{\exp(-z)}{1+\exp(-z)} = -\frac{1}{1+\exp(z)}$

↪ The second derivative $L^{''}(z) = \frac{\partial L^{'}}{\partial z} = +\frac{\exp(z)}{(1+\exp(z))^2} > 0$

↪ Hence, the logistic loss $L(z)$ is a convex function.

5 Programming: Recommendation System [40 pts]

Personalized recommendation systems are used in a wide variety of applications such as electronic commerce, social networks, web search, and more. Machine learning techniques play a key role to extract individual preference over items. In this assignment, we explore this popular business application of machine learning, by implementing a simple matrix-factorization-based recommender using gradient descent.

Suppose you are an employee in Netflix. You are given a set of ratings (from one star to five stars) from users on many movies they have seen. Using this information, your job is implementing a personalized rating predictor for a given user on unseen movies. That is, a rating predictor can be seen as a function $f : \mathcal{U} \times \mathcal{I} \rightarrow \mathbb{R}$, where $\mathcal{U}$ and $\mathcal{I}$ are the set of users and items, respectively. Typically the range of this function is restricted to between 1 and 5 (stars), which is the the allowed range of the input.

Now, let's think about the data representation. Suppose we have m users and n items, and a rating given by a user on a movie. We can represent this information as a form of matrix, namely rating matrix M. Suppose rows of M represent users, while columns do movies. Then, the size of matrix will be $m \times n$. Each cell of the matrix may contain a rating on a movie by a user. In $M_{15,47}$, for example, it may contain a rating on the item 47 by user 15. If he gave 4 stars, $M_{15,47} = 4$. However, as it is almost impossible for everyone to watch large portion of movies in the market, this rating matrix should be very sparse in nature. Typically, only 1% of the cells in the rating matrix are observed in average. All other 99% are missing values, which

means the corresponding user did not see (or just did not provide the rating for) the corresponding movie. Our goal with the rating predictor is estimating those missing values, reflecting the user's preference learned from available ratings.

Our approach for this problem is matrix factorization. Specifically, we assume that the rating matrix M is a low-rank matrix. Intuitively, this reflects our assumption that there is only a small number of factors (e.g, genre, director, main actor/actress, released year, etc.) that determine like or dislike. Let's define r as the number of factors. Then, we learn a user profile $U \in \mathbb{R}^{m \times r}$ and an item profile $V \in \mathbb{R}^{n \times r}$. (Recall that m and n are the number of users and items, respectively.) We want to approximate a rating by an inner product of two length r vectors, one representing user profile and the other item profile. Mathematically, a rating by user u on movie i is approximated by

$$M_{u,i} \approx \sum_{k=1}^{r} U_{u,k} V_{i,k}. \tag{3}$$

We want to fit each element of U and V by minimizing squared reconstruction error over all training data points. That is, the objective function we minimize is given by

$$E(U,V) = \sum_{(u,i) \in M} (M_{u,i} - U_u^T V_i)^2 = \sum_{(u,i) \in M} (M_{u,i} - \sum_{k=1}^{r} U_{u,k} V_{i,k})^2 \tag{4}$$

where U_u is the uth row of U and V_i is the ith row of V. We observe that this looks very similar to the linear regression. Recall that we minimize in linear regression:

$$E(\theta) = \sum_{i=1}^{m} (Y^i - \theta^T x^i)^2 = \sum_{i=1}^{m} (Y^i - \sum_{k=1}^{r} \theta_k x_k^i)^2 \tag{5}$$

where m is the number of training data points. Let's compare (4) and (5). $M_{u,i}$ in (4) corresponds to Y^i in (5), in that both are the observed labels. $U_u^T V_i$ in (4) corresponds to $\theta^T x^i$ in (5), in that both are our estimation with our model. The only difference is that both U and V are the parameters to be learned in (4), while only θ is learned in (5). This is where we personalize our estimation: with linear regression, we apply the same θ to any input x^i, but with matrix factorization, a different profile U_u are applied depending on who is the user u.

As U and V are interrelated in (4), there is no closed form solution, unlike linear regression case. Thus, we need to use gradient descent:

$$U_{v,k} \leftarrow U_{v,k} - \mu \frac{\partial E(U,V)}{\partial U_{v,k}}, \qquad V_{j,k} \leftarrow V_{j,k} - \mu \frac{\partial E(U,V)}{\partial V_{j,k}}, \tag{6}$$

where μ is a hyper-parameter deciding the update rate. It would be straightforward to take partial derivatives of $E(U,V)$ in (4) with respect to each element $U_{v,k}$ and $V_{j,k}$. Then, we update each element of U and V using the gradient descent formula in (6).

(a) Derive the update formula in (6) **by solving the partial derivatives. [10 pts]**

$\hookrightarrow$ We have $\dfrac{\partial E(U,V)}{\partial U_{v,k}} = \dfrac{\partial \sum_j (M_{v,j} - \sum_{k=1}^{r} U_{v,k} V_{j,k})^2}{\partial U_{v,k}}$

$\hookrightarrow$ Thus, $\dfrac{\partial E(U,V)}{\partial U_{v,k}} = -2 \sum_j V_{j,k} (M_{v,j} - \sum_{k=1}^{r} U_{v,k} V_{j,k})$

$\hookrightarrow$ We have $\dfrac{\partial E(U,V)}{\partial V_{j,k}} = \dfrac{\partial \sum_v (M_{v,j} - \sum_{k=1}^{r} U_{v,k} V_{j,k})^2}{\partial V_{j,k}}$

$\hookrightarrow$ Thus, $\frac{\partial E(U,V)}{\partial V_{j,k}} = -2\sum_v U_{v,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k})$

$\hookrightarrow$ Hence, the formula in (6) are updated as below

- $U_{v,k} \leftarrow U_{v,k} + 2\mu\sum_j V_{j,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k})$
- $V_{j,k} \leftarrow V_{j,k} + 2\mu\sum_v U_{v,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k})$

(b) To avoid overfitting, we usually add regularization terms, which penalize for large values in U and V. Redo part (a) using the regularized objective function below. [5 pts]

$$E(U,V) = \sum_{(u,i)\in M} (M_{u,i} - \sum_{k=1}^{r} U_{u,k}V_{i,k})^2 + \lambda\sum_{u,k} U_{u,k}^2 + \lambda\sum_{i,k} V_{i,k}^2$$

(λ is a hyper-parameter controlling the degree of penalization.)

- To avoid over-fitting, we can add the derivative part of the regularization to the previous results to get the required equations. See below :

$\hookrightarrow$ $\frac{\partial E(U,V)}{\partial U_{v,k}} = -2\sum_j V_{j,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k}) + 2\lambda U_{v,k}$

$\hookrightarrow$ $\frac{\partial E(U,V)}{\partial V_{j,k}} = -2\sum_v U_{v,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k}) + 2\lambda V_{j,k}$

- Thus, the update formulae are as below (Like (6)):

$\hookrightarrow$ $U_{v,k} \leftarrow U_{v,k} + 2\mu\sum_j V_{j,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k}) - 2\lambda\mu U_{v,k}$

$\hookrightarrow$ $V_{j,k} \leftarrow V_{j,k} + 2\mu\sum_v U_{v,k}(M_{v,j} - \sum_{k=1}^{r} U_{v,k}V_{j,k}) - 2\lambda\mu V_{j,k}$

(c) Implement `myRecommender.m/myRecommender.py` by filling the gradient descent part.

You are given a skeleton code `myRecommender.m/myRecommender.py`. Using the training data `rateMatrix`, you will implement your own recommendation system of rank `lowRank`. The only file you need to edit and submit is `myRecommender.m`. In the gradient descent part, repeat your update formula in (b), observing the average reconstruction error between your estimation and ground truth in training set. You need to set a stopping criteria, based on this reconstruction error as well as the maximum number of iterations. You should play with several different values for μ and λ to make sure that your final prediction is accurate.

Formatting information is here:

Input

- **rateMatrix**: training data set. Each row represents a user, while each column an item. Observed values are one of $\{1, 2, 3, 4, 5\}$, and missing values are 0.
- **lowRank**: the number of factors (dimension) of your model. With higher values, you would expect more accurate prediction.

Output

- **U**: the user profile matrix of dimension user count $\times$ low rank.
- **V**: the item profile matrix of dimension item count $\times$ low rank.

Evaluation [15 pts]

Upload your `myRecommender.m/myRecommender.py` implementation file. (Do not copy and paste your code in your report. Be sure to upload your `myRecommender.m/myRecommender.py` file.)

To test your code, try to run `homework3.m/homework3.py`. You may have noticed that the code prints both training and test error, in RMSE (Root Mean Squared Error), defined as follows:

$$\sum_{(u,i)\in M} (M_{u,i} - f(u,i))^2$$

where $f(u,i)$ is your estimation, and the summation is over the training set or testing set, respectively. For the grading, we will use another set-aside testing set, which is not released to you. If you observe your test error is less than 1.00 without cheating (that is, training on the test set), you may expect to see the similar performance on the unseen test set as well.

Note that we provide `homework3.m/homework3.py` just to help you evaluate your code easily. You are not expected to alter or submit this to us. In other words, we will not use this file when we grade your submission. The only file we take care of is `myRecommender3.m/myRecommender3.py`.

Grading criteria:

- Your code should output U and V as specified. The dimension should match to the specification.
- We will test your output on another test dataset, which was not provided to you. The test RMSE on this dataset should be at least 1.05 to get at least partial credit.
- We will measure elapsed time for learning. If your implementation takes longer than 3 minutes for rank 5, you should definitely try to make your code faster or adjust parameters. Any code running more than 5 minutes is not eligible for credit.
- Your code should not crash. Any crashing code will be not credited.

Report [10 pts]

In your report, show the performance (RMSE) both on your training set and test set, with varied `lowRank`. (The default is set to 1, 3, and 5, but you may want to vary it further.) Discuss what you observe with varied low rank. Also, briefly discuss how you decided your hyper-parameters (μ, λ).

Note

- Do not print anything in your code (e.g, iteration 1 : err=2.4382) in your final submission.
- Do not alter input and output format of the skeleton file. (E.g, adding a new parameter without specifying its defalut value) Please make sure that you returned all necessary outputs according to the given skeleton.
- Please do not use additional file. This task is simple enough that you can fit in just one file.
- Submit your code with the best parameters you found. We will grade without modifying your code. (Applying cross-validation to find best parameters is fine, though you do not required to do.)
- Please be sure that your program finishes within a fixed number of iterations. Always think of a case where your stopping criteria is not satisfied forever. This can happen anytime depending on the data, not because your code is incorrect. For this, we recommend setting a maximum number of iteration in addition to other stopping criteria.

Grand Prize

Similar to the Netflix competition held in 2006 to 2009, the student who achives the lowest RMSE on the secret test set will earn the Grand Prize. We will give extra 10 bonus points to the winner, and share the student's code to everyone. (Note that the winner should satisfy all other grading criteria perfectly, including answer sanity check and timing requirement. Otherwise, the next student will be considered as the winner.)

Typing Bonus

As usual, we will give 5 bonus points for typed submissions. Note that **all** questions should be typed to get this credit.

lowRank	Train RMSE	Test RMSE	Run-time
1	0.9162	0.9471	16.86
3	0.8587	0.9358	16.41
5	0.8401	0.9427	17.10
7	0.8160	0.9424	18.12
9	0.7945	0.9370	19.14

Figure 1: The output of the experiment (Run-time is in seconds)

↪ From the table, we see that the RMSE of train set decreases with increase in `lowRank`. This is intuitive as higher `lowRank` implies more information given to the model, and hence better predictions. This also means higher run-time as seen in the table.

↪ For the test set, we see that it varies non-uniformly, but it stays around the range of [0.9370, 0.9471].

- **Hyper-parameters are decided as below :**

↪ To decide values of μ and λ, it is more of a grid-search type approach (since this is usually varying on the basis of the data-set).

↪ For a value of the learning constant μ, too big of a value will make the solution over-shoot. And too small of μ would require an excessive amount of time to converge.

↪ The values used are $\mu = 0.0002$ (learning_rate) and $\lambda = 0.02$ (reg_coef), with max_iter=500 (maximum iteration count) - determined through grid-search (trial-error).