Algorithm Implementation & Evaluation

Project 4 - Group 3

Adam Kravitz
Sungin Cho
Luyue Chen
Xiaotong Li
Syed Ahsan Bukhari

Goal

A1 vs A2 Given P3

A1 - Stochastic Gradient Descent

A2 - Probabilistic Matrix Factorization

P3 - Post Processing SVD with Kernel Ridge Regression

Project Workflow

Step 1

Hard-code A1, A2 and P3.

Step 2

Parameters Tuning to A1 and A2 separately by Cross Validation.

Step 3

Use the best parameters to train the model.

Step 4

Apply P3 to A1, A2
Perform Evaluation

A1 - Stochastic Gradient Descent

- Minimum possible value
- Sample to run each iteration
- Efficient as compared to batch gradient
- Smaller steps (learning rate)

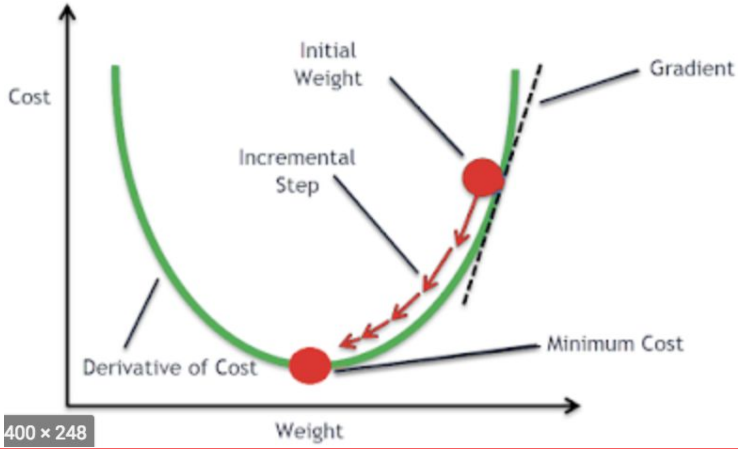
$$min_{q^*p^*} \sum_{(u,i)\in K} (r_{ui} - q_i^T p_u)^2 + \lambda(||q_i||^2 + ||p_u||^2)$$

$$e_{ui} = r_{ui} - q_i^T p$$

- Partial Derivative

$$q_i = q_i + g(e_{ui} \cdot p_u - \lambda \cdot q_i)$$

$$p_u = p_u + g(e_{ui} \cdot q_i - \lambda \cdot p_u)$$



A1 - Stochastic Gradient Descent

Tuning Parameters

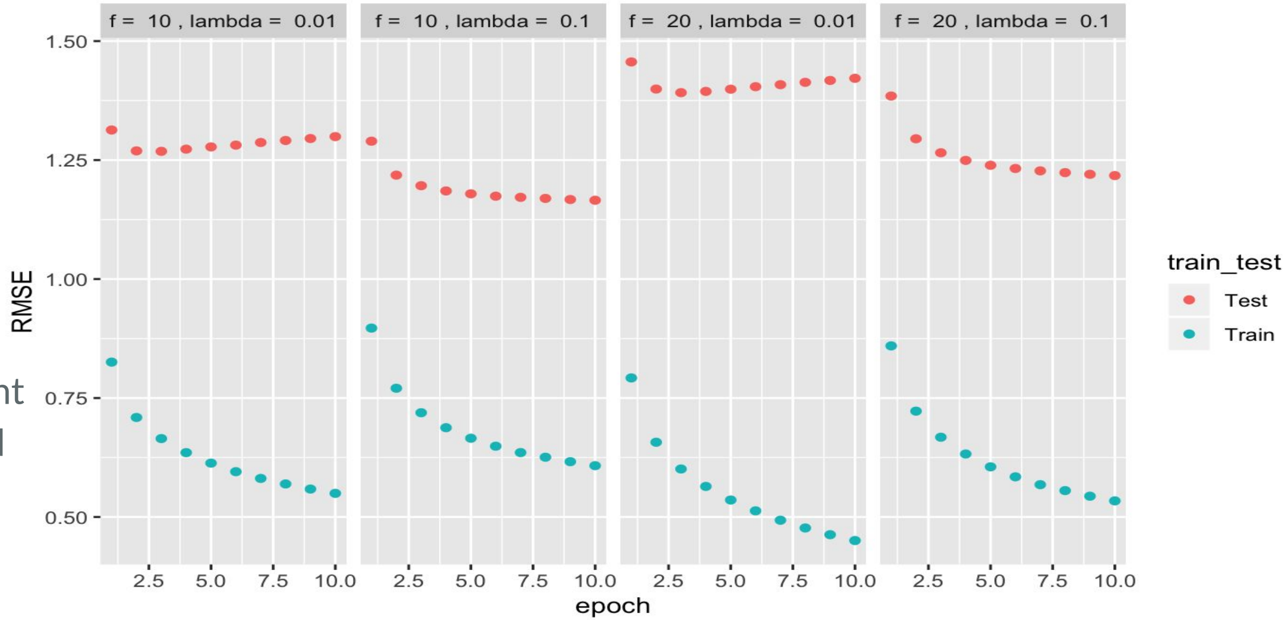
F = 10

Lambda = 0.1

Train RMSE: 0.5675

Test RMSE: 1.1673

Tuned by combinations of different lambdas and f, evaluated by speed and criterion.



A2 - Probabilistic Gradient Descent

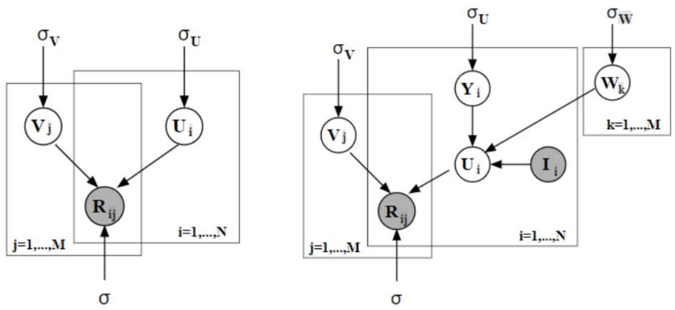
$$E = \frac{1}{2}\sum_{i=1}^{M}\sum_{u=1}^{U} I_{iu}(r_{ui} - q_i^T p_u)^2 + \frac{\sigma}{2\sigma_q}\sum_{i=1}^{M} ||q_i||^2 + \frac{\sigma}{2\sigma_p}\sum_{u=1}^{U} ||p_u||^2$$

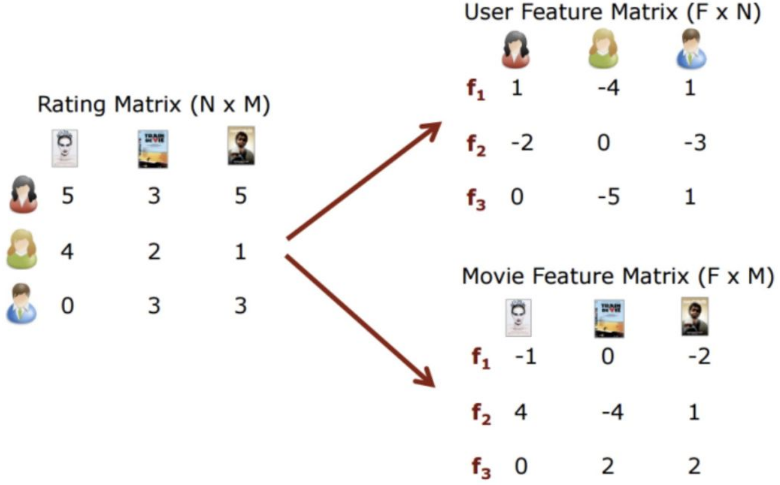
$$p_u = p_u + lrate * (-Grad_p)$$

$$q_i = q_i + lrate * (-Grad_q)$$

$$Grad_p = \frac{\partial L}{\partial p_u} = -\sum_{i=1}^{I} I_{ui}(r_{ui} - p_u^t q_i) q_i + \lambda_p p_u$$

$$Grad_q = \frac{\partial L}{\partial q_i} = -\sum_{u=1}^{U} I_{ui}(r_{ui} - p_u^t q_i) p_u + \lambda_q q_i$$





I_{iu} = Is an identifier, saying if a user rated some movie its a 1 otherwise it's a 0.

A2 - Probabilistic Matrix Factorization

Tuning Parameters

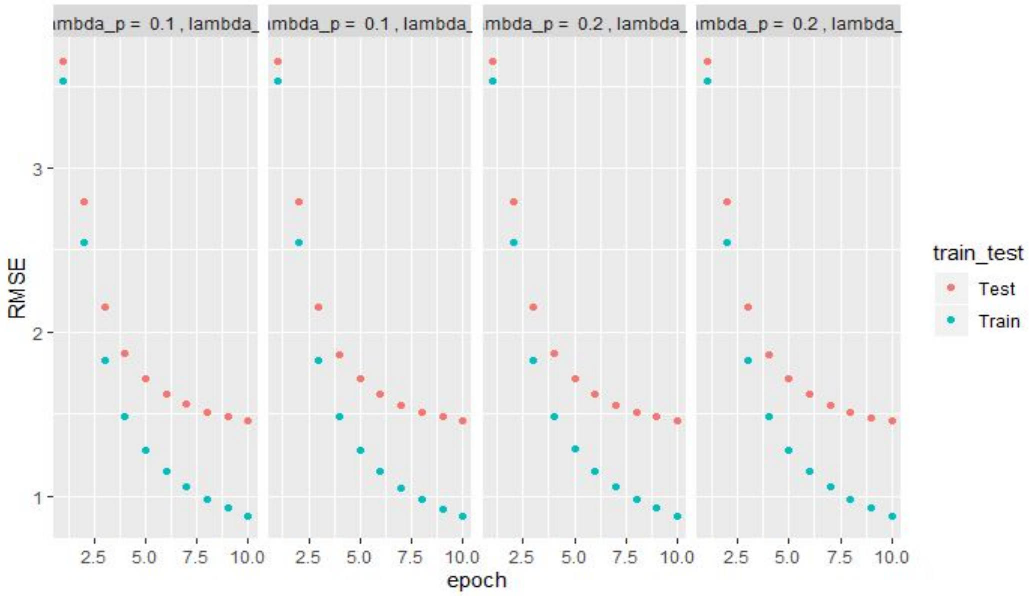
F= 8,

Lambda_p = 0.1,

Lambda_q = 0.1

lr=0.0005

Tuned by combinations of different lambdas and f, evaluated by speed and criterion.



P3 - SVD with Kernel Ridge Regression

Improving SVD by discarding weights of user ($p_u\ from\ r = q_i^t p_u\ in\ A1\ or\ A2$) after training and predict y_{ij} for each user i using ridge regression.

Steps

1. Make X

- each row is normalized vector of features of one movie j rated by user i. $x_j = \frac{q_j}{\|q_j\|}$.

```
# Making X, predictor matrix
X_mat <- function(q){
  V <- t(q)
  X <- apply(V, 2, function(x) x/Norm(x)) # make X using v_j (mxk)
  return(X)
}
```

2. Make function $K(X,X)$, Gaussian kernel

- $K(x_i^T, x_j^T) = exp(2(x_i^T x_j - 1))$

```
# Gaussian kernel (t_x_i: transpose of x_i)
K <- function(t_x_i, X){
  return(exp(2*(t_x_i%*%t(X))-1))
}
```

3. Get new prediction of rating after postprocessing using the prediction model

- $\hat{y}_i = K(x_i^T, X)(K(X,X) + \lambda I)^{-1} y$:

$y\ is\ the\ prediction\ rating\ from\ the\ result\ of\ A1\ or\ A2\ /\ \lambda = 0.5, used\ same\ value\ for\ \lambda\ in\ the\ paper.$

```
# SVD with kernel ridge regression
svd_krr <- function(n, lamb=0.5, X, pred_rate){
  I <- diag(n)
  return(K(X,X) %*% solve((K(X,X)+lamb*I)) %*% pred_rating)
}
```

A1+P3 Results

Stochastic Gradient Descent post processed by SVD with Ridge Kernel Regression

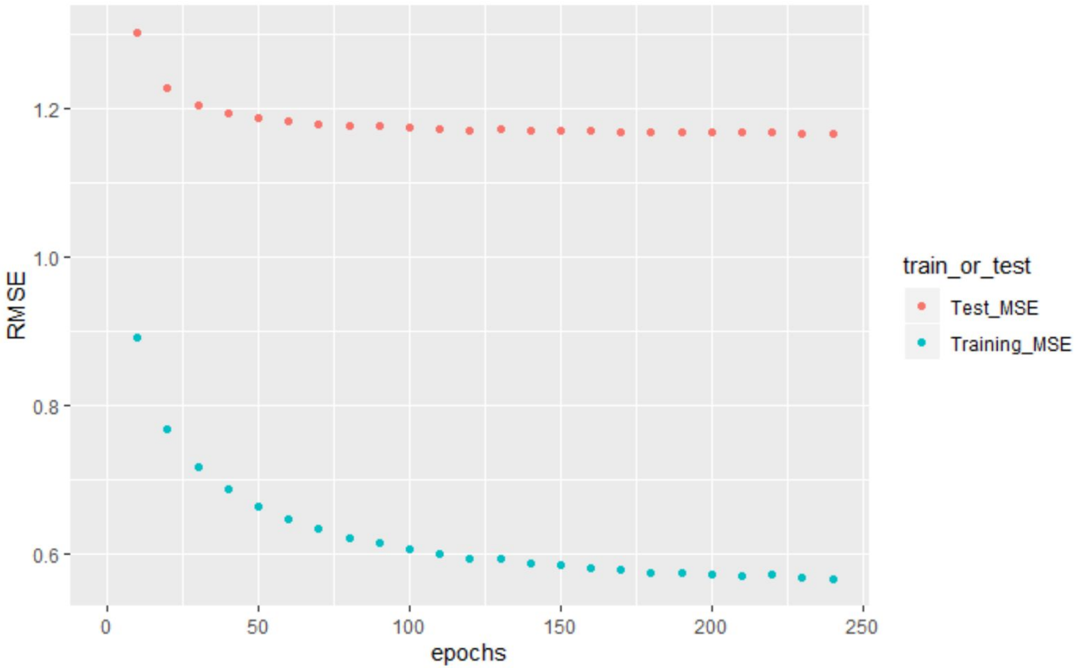
# of Epoch	Training RMSE	Testing RMSE	Testing RMSE After P3	Processing time
240	0.5675	1.167317	0.6954	86:46:29

Parameters: f = 10, lr = 0.01, lambda= 0.1

- f is the dimension of the decomposed vectors
- lr is the learning rate for each gradient descent steps.
- lambda is the coefficient on the regularized term.

Evaluation (A1+P3)

- The gap between Testing RMSE and Training RMSE is large, but become stable after 50 epochs.
- The Test RMSE at 50 epochs is around 1.1871
- The running time for the first 10 epochs is around 14 mins and the time after the first epoch drastically decreased to average 2mins.



A2+P3 Results

Gradient Descent Post processed by SVD with Ridge Kernel Regression

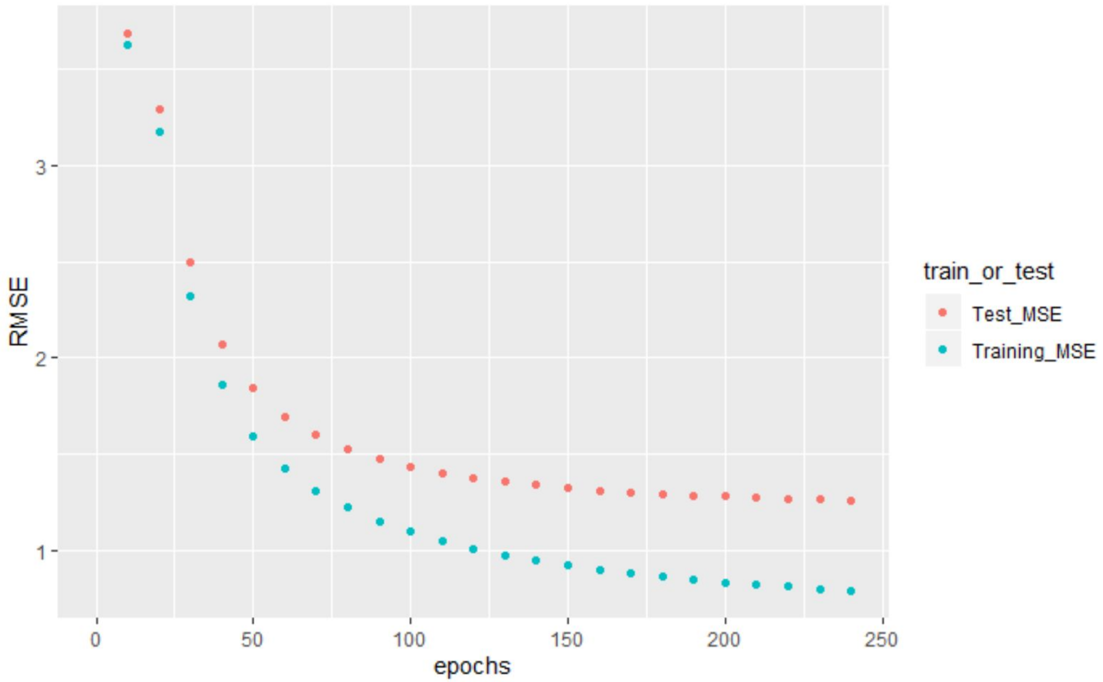
# of Epoch	Training RMSE	Testing RMSE	Testing RMSE after P3	Processing time
240	0.7955	1.2641	**0.8659**	61:54:33

Parameters: f = 8, lr = 0.0005, lambda_p = 0.1, lambda_q=0.1

- f is the dimension of the decomposed vectors
- lr is the learning rate for each gradient descent steps.
- lambda_p, lambda_q are the coefficients on the regularized terms.

Evaluation(A2+P3)

- Gap between training RMSE and testing RMSE get wider → Signs of over-fitting
- The testing RMSE starting to become stable after 150 epochs. We should stopped the training process from there.
- Test RMSE at 150 epoch = 1.326
- After Post Processing, the RMSE dropped to 0.8659



Result

Accuracy: A1 + P3 has better RMSE as compared to A2 + P3

Computation Time: A2 + P3 has better RMSE as compared to A1 + P3