Mohit Kumar
SR No. 04-01-03-10-51-21-1-19825
MTech Artificial Intelligence

Assignment 6: Recommendation Systems

---

python packages used:

- numpy
- pandas
- scipy
- matplotlib

Implementation details I have implemented the following functions:

plot_error_SVD This function takes as input a matrix in sparse format i.e. COOrdinate format of sparse module of scipy and a range for latent factors and generates the plots of mean square error vs number of latent factors, the computation time vs number of latent factors and storage used vs number of latent factors.

sample_k This function takes as input a sparse matrix and an integer k and returns a sparse matrix made up of k randomly selected rows and columns according to the energy distribution of each row and column.

plot_error_CUR This function takes as input a matrix in sparse format i.e. COOrdinate format of sparse module of scipy and a range for latent factors and generates the plots of mean square error vs number of latent factors, the computation time vs number of latent factors and storage used vs number of latent factors.

get_matrix This function takes as input one subset dataframe and the full dataframe and returns the numpy matrix corresponding to the subset.

train_test_split This function takes as input the full dataframe and the test_ratio and returns the numpy matrices corresponding to the train and test splits.

get_gradient_P This function returns the gradient of the loss with respect to P.

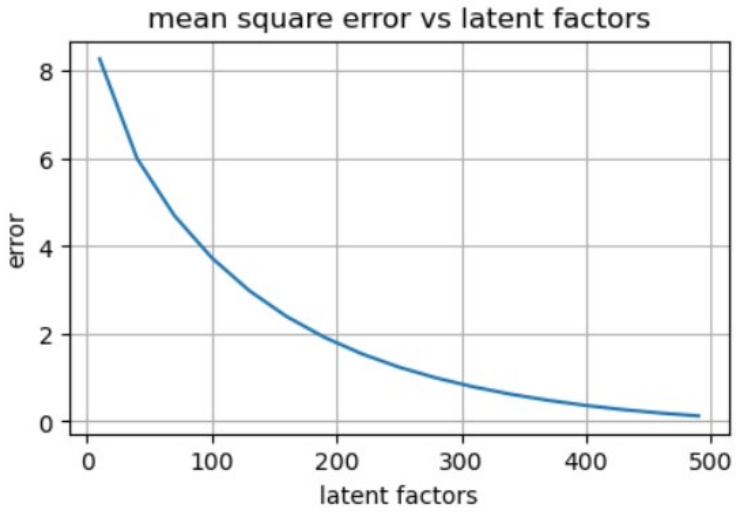
get_gradient_Q This function returns the gradient of the loss with respect to Q.

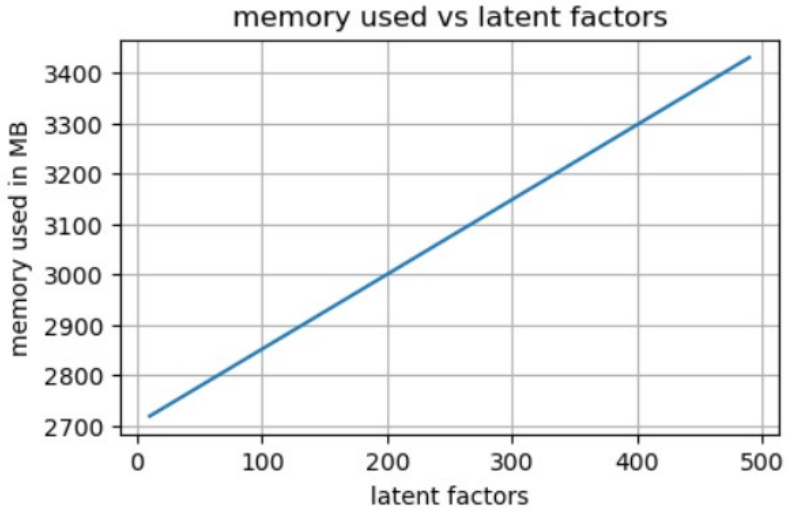
get_loss This function returns the loss function value.

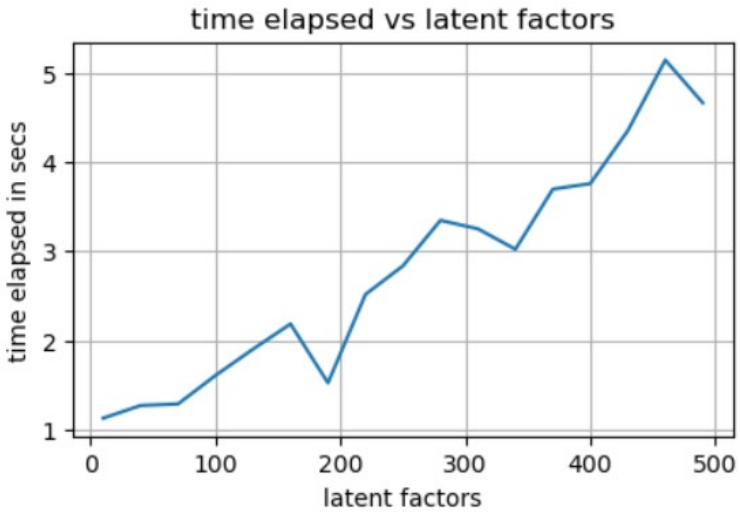
gradient_descent This function performs the gradient descent to find the best P and Q for a given matrix.

Results for the small dataset

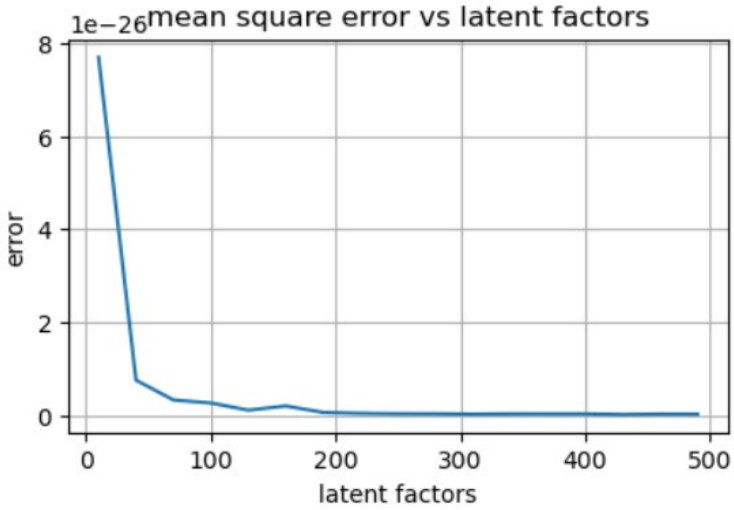
Singular Value Decomposition

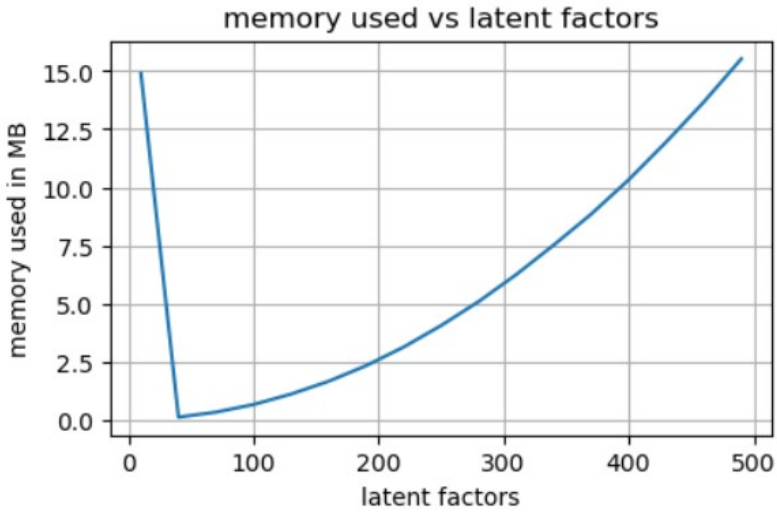


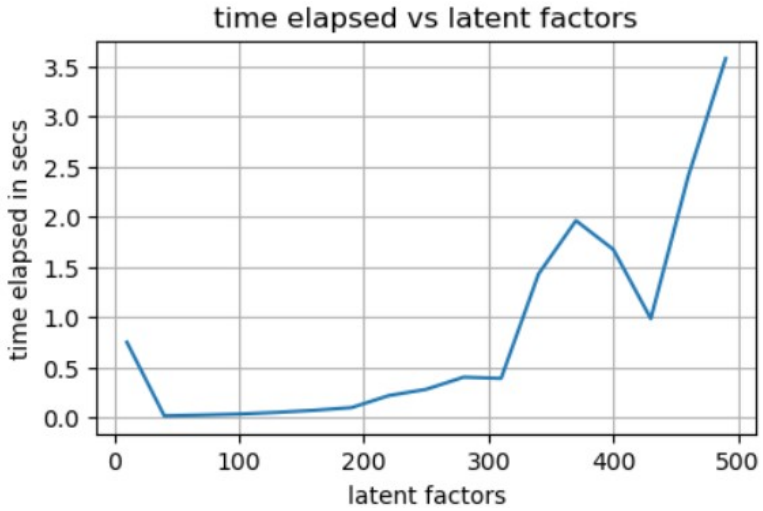




CUR Decomposition



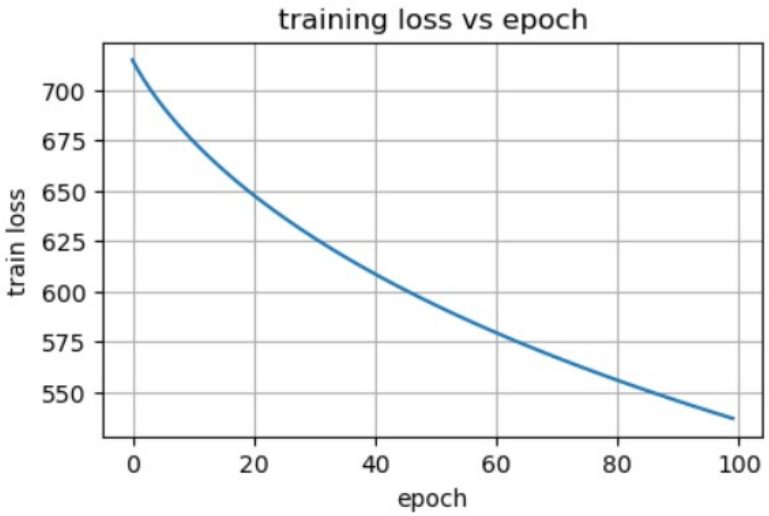




PQ decomposition

hyperparameters used

- number of latent factors: 20
- regularization(lambda value): 0.01
- learning rate: 0.01
- train-test split: 80:20
- stochastic gradient descent
- epochs: 100



Execution time for PQ decomposition: 68 minutes
Final test mean square error for PQ decomposition: 3.59

Observations

Question 1 SVD was applied on the small dataset. We observe that to capture most of the information more than 400 latent factors are required. We see from the MSE plot for the SVD that the mean square error is significantly reduced if we use more than 400 latent factors.

Question 2 CUR Decomposition was applied on the small dataset. For a given number of latent factors, the additional error introduced by CUR decomposition keeps varying as compared to the SVD decomposition. This is because we are randomly selecting rows and columns.

Question 3 Running time of the CUR decomposition is lesser compared to the SVD. Storage requirement is also less in the case of the CUR decomposition as can be observed from the plots.