Homework 2: Machine Learning

1) Collaborative Filtering:

Steps:

1. Converted text file to csv files with appropriate column names.
2. Counted number of unique users and movies in dataset.
3. Create a pivot table of training and testing datasets and replace every unavailable data with a zero.
4. Convert the pivot tables into multidimensional array.
5. Assign weights to parameters.
6. Predict using the formula given in the Research paper.
7. Compare the predicted value with the test value using Mean Square Error method.

2) Neural Networks, K-nearest neighbors and SVMs:

i) SVM

SVM classification was performed using svm.SVC library of sklearn.

Different parameters used in SVC were:

- Regularization parameter C is set to different values from 1 to 500.
- All the kernels were used with different parameter settings.
- Degree of poly kernel is set to different values.
- Max iteration is also used to stop early.

Best error rate in case of SVM was achieved using rbf kernel with regularization parameter C = 500 and gamma = 0.05.

ii) MLPClassifier:

MLP Classification is performed using MLPClassifier library of sklearn.

Different parameters used in SVC were:

- Different size of hidden layers is used from 50 to 250
- All the solver for weight optimization is used with different parameters.
- Alpha penalty is set from 0.01 to 0.001
- Batch size for lbfgs solver is set to different values.

- Beta1 is varied from 09 to 0.95 and beta2 is varied from 09 to 0.999.
- Max epochs are set from 10 to 20.

Best error rate in case of MLP was achieved using lbfgs and adam solvers for weight optimization. The number of layers used were between 200 and 250 and learning rate was set to invscaling.

iii) K Nearest Neighbors:

K Nearest Neighbors is performed using KNeighborsClassifier library of sklearn.

Different parameters used in KNeighborsClassifier were:

- Classification is performed using 3, 5 and 7 neighbors.
- Size of leaf nodes varies from 10 to 30.
- All combination of algorithm, metric and number of neighbors are used to compare the outputs.

Best error rate in case of KNeighborsClassifier was achieved using 3 nearest neighbors and when weight function, algorithm, leaf size, and metric were set to distance, ball tree, 10 and Euclidean respectively.

A similar error rate was achieved using same parameters but with 7 nearest neighbors.