Machine Learning: Programming Exercise 8

**Anomaly Detection and Recommender Systems**

In this exercise, you will implement the anomaly detection algorithm and apply it to detect failing servers on a network. In the second part, you will use collaborative filtering to build a recommender system for movies.

**Files needed for this exercise**

* ex8.mlx - MATLAB Live Script that steps you through the exercise
* ex8data1.mat - First example Dataset for anomaly detection
* ex8data2.mat - Second example Dataset for anomaly detection
* ex8\_movies.mat - Movie Review Dataset
* ex8\_movieParams.mat - Parameters provided for debugging
* multivariateGaussian.m - Computes the probability density function for a Gaussian distribution
* visualizeFit.m - 2D plot of a Gaussian distribution and a dataset
* checkCostFunction.m - Gradient checking for collaborative filtering
* computeNumericalGradient.m - Numerically compute gradients
* fmincg.m - Function minimization routine (similar to fminunc)
* loadMovieList.m - Loads the list of movies into a cell-array
* movie\_ids.txt - List of movies
* normalizeRatings.m - Mean normalization for collaborative filtering
* submit.m - Submission script that sends your solutions to our servers
* \*estimateGaussian.m - Estimate the parameters of a Gaussian distribution with a diagonal covariance matrix
* \*selectThreshold.m - Find a threshold for anomaly detection
* \*cofiCostFunc.m - Implement the cost function for collaborative filtering

***\* indicates files you will need to complete***

**Clear existing variables and confirm that your Current Folder is set correctly**

Click into this section, then click the 'Run Section' button above. This will execute the clear command to clear existing variables and the dir command to list the files in your Current Folder. The output should contain all of the files listed above and the 'lib' folder. If it does not, right-click the 'ex8' folder and select 'Open' before proceding or see the instructions in README.mlx for more details.

clear

dir

**Before you begin**

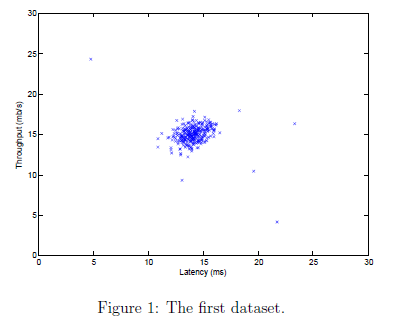
The workflow for completing and submitting the programming exercises in MATLAB Online differs from the original course instructions. Before beginning this exercise, make sure you have read through the instructions in README.mlx which is included with the programming exercise files. README also contains solutions to the many common issues you may encounter while completing and submitting the exercises in MATLAB Online. Make sure you are following instructions in README and have checked for an existing solution before seeking help on the discussion forums.

**Table of Contents**

[Anomaly Detection and Recommender Systems](about:blank#H_47AD2164)  [Files needed for this exercise](about:blank#H_19F66266)  [Clear existing variables and confirm that your Current Folder is set correctly](about:blank#H_2D10AFC9)  [Before you begin](about:blank#H_9BD8929B)[1. Anomaly Detection](about:blank#H_9038F655)  [1.1 Gaussian distribution](about:blank#H_21766FD3)  [1.2 Estimating parameters for a Gaussian](about:blank#H_FEE8E6EE)  [1.3 Selecting the threshold,](about:blank#H_B710CE62)  [1.4 High dimensional dataset](about:blank#H_CB55AD31)[2. Recommender Systems](about:blank#H_70E8E5A1)  [2.1 Movie ratings dataset](about:blank#H_737015E6)  [2.2 Collaborative filtering learning algorithm](about:blank#H_A796401B)  [2.2.1 Collaborative filtering cost function](about:blank#H_92F1A768)  [2.2.2 Collaborative filtering gradient](about:blank#H_981587CE)  [2.2.3 Regularized cost function](about:blank#H_500A8E84)  [2.2.4 Regularized gradient](about:blank#H_65140049)  [2.3 Learning movie recommendations](about:blank#H_37136E48)  [2.3.1 Recommendations](about:blank#H_035E8F2B)[Submission and Grading](about:blank#H_A55977AC)

**1. Anomaly Detection**

In this exercise, you will implement an anomaly detection algorithm to detect anomalous behavior in server computers. The features measure the throughput (mb/s) and latency (ms) of response of each server. While your servers were operating, you collected  examples of how they were behaving, and thus have an unlabeled dataset . You suspect that the vast majority of these examples are 'normal' (non-anomalous) examples of the servers operating normally, but there might also be some examples of servers acting anomalously within this dataset.



You will use a Gaussian model to detect anomalous examples in your dataset. You will first start on a 2D dataset that will allow you to visualize what the algorithm is doing. On that dataset you will fit a Gaussian distribution and then find values that have very low probability and hence can be considered anomalies. After that, you will apply the anomaly detection algorithm to a larger dataset with many dimensions. The code below will visualize the dataset as shown in Figure 1.

% The following command loads the dataset. You should now have the variables X, Xval, yval in your environment

load('ex8data1.mat');

% Visualize the example dataset

plot(X(:, 1), X(:, 2), 'bx');

axis([0 30 0 30]);

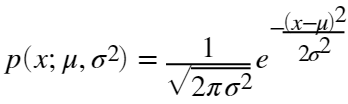
xlabel('Latency (ms)');

ylabel('Throughput (mb/s)');

**1.1 Gaussian distribution**

To perform anomaly detection, you will first need to fit a model to the data's distribution. Given a training set  (where ), you want to estimate the Gaussian distribution for each of the features . For each feature  you need to find parameters  and  that fit the data in the *i*-th dimension (the *i*-th dimension of each example).

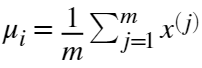
The Gaussian distribution is given by



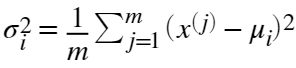
where *μ* is the mean and  controls the variance.

**1.2 Estimating parameters for a Gaussian**

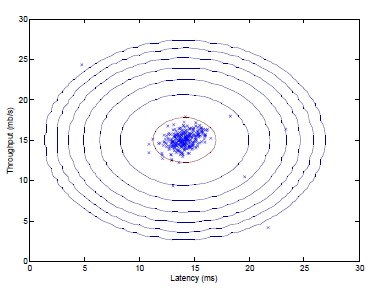
You can estimate the parameters, , of the *i*-th feature by using the following equations. To estimate the mean, you will use:

,

and for the variance you will use:

.

Your task is to complete the code in estimateGaussian.m. This function takes as input the data matrix X and should output an *n*-dimension vector mu that holds the mean of all the *n* features and another *n*-dimension vector sigma2 that holds the variances of all the features. You can implement this using a for loop over every feature and every training example (though a vectorized implementation might be more efficient; feel free to use a vectorized implementation if you prefer). Note that in MATLAB, the var function will (by default) use , instead of  when computing .



Once you have completed the code in estimateGaussian.m, the code below will visualize the contours of the fitted Gaussian distribution. You should get a plot similar to Figure 2. From your plot, you can see that most of the examples are in the region with the highest probability, while the anomalous examples are in the regions with lower probabilities.

% Estimate mu and sigma2

[mu, sigma2] = estimateGaussian(X);

% Returns the density of the multivariate normal at each data point (row) of X

p = multivariateGaussian(X, mu, sigma2);

% Visualize the fit

visualizeFit(X, mu, sigma2);

xlabel('Latency (ms)');

ylabel('Throughput (mb/s)');

*You should now submit your solutions. Enter* **submit***at the command prompt, then and enter or confirm your login and token when prompted.*

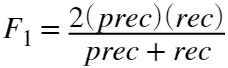
**1.3 Selecting the threshold,** *ϵ*

Now that you have estimated the Gaussian parameters, you can investigate which examples have a very high probability given this distribution and which examples have a very low probability. The low probability examples are more likely to be the anomalies in our dataset. One way to determine which examples are anomalies is to select a threshold based on a cross validation set. In this part of the exercise, you will implement an algorithm to select the threshold *ϵ* using the  score on a cross validation set.

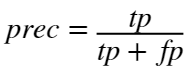
You should now complete the code in selectThreshold.m. For this, we will use a cross validation set , where the label  corresponds to an anomalous example, and  corresponds to a normal example. For each cross validation example, we will compute . The vector of all of these probabilities  is passed to selectThreshold.m in the vector pval. The corresponding labels  is passed to the same function in the vector yval.

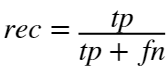
The function selectThreshold.m should return two values; the first is the selected threshold *ϵ*. If an example *x* has a low probability , then it is considered to be an anomaly. The function should also return the  score, which tells you how well you're doing on finding the ground truth anomalies given a certain threshold. For many different values of *ϵ*, you will compute the resulting  score by computing how many examples the current threshold classifies correctly and incorrectly.

The  score is computed using precision (*prec*) and recall (*rec*):

,

You compute precision and recall by:

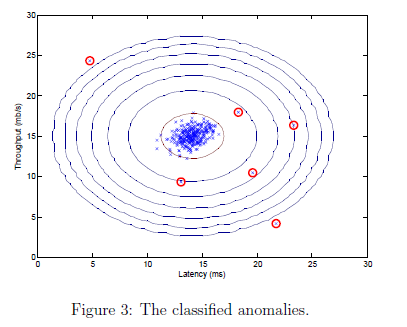




where

*  is the number of true positives: the ground truth label says it's an anomaly and our algorithm correctly classified it as an anomaly.
*  is the number of false positives: the ground truth label says it's not an anomaly, but our algorithm incorrectly classified it as an anomaly.
*  is the number of false negatives: the ground truth label says it's an anomaly, but our algorithm incorrectly classified it as not being anomalous.

In the provided code selectThreshold.m, there is already a loop that will try many different values of *ϵ* and select the "best" based *ϵ* on the  score. You should now complete the code in selectThreshold.m. You can implement the computation of the  score using a for loop over all the cross validation examples (to compute the values . You should see a value for epsilon of about 8.99e-05.



**Implementation Note**: In order to compute , you may be able to use a vectorized implementation rather than loop over all the examples. This can be implemented by MATLAB's equality test between a vector and a single number. If you have several binary values in an *n*-dimensional binary vector , you can find out how many values in this vector are 0 by using: sum(v == 0). You can also apply a logical and operator to such binary vectors. For instance, let cvPredictions be a binary vector of the size of your cross validation set, where the *i*-th element is 1 if your algorithm considers  an anomaly, and 0 otherwise. You can then, for example, compute the number of false positives using: fp = sum((cvPredictions == 1) & (yval == 0)).

Once you have completed the code in selectThreshold.m, the code below will detect and circle the anomalies in the plot (Figure 3).

pval = multivariateGaussian(Xval, mu, sigma2);

[epsilon, F1] = selectThreshold(yval, pval);

fprintf('Best epsilon found using cross-validation: %e\n', epsilon);

fprintf('Best F1 on Cross Validation Set: %f\n', F1);

% Find the outliers in the training set and plot the

outliers = find(p < epsilon);

% Visualize the fit

visualizeFit(X, mu, sigma2);

xlabel('Latency (ms)');

ylabel('Throughput (mb/s)');

% Draw a red circle around those outliers

hold on

plot(X(outliers, 1), X(outliers, 2), 'ro', 'LineWidth', 2, 'MarkerSize', 10);

hold off

*You should now submit your solutions. Enter* **submit***at the command prompt, then and enter or confirm your login and token when prompted.*

**1.4 High dimensional dataset**

The code in this section will run the anomaly detection algorithm you implemented on a more realistic and much harder dataset. In this dataset, each example is described by 11 features, capturing many more properties of your compute servers. The code below will use your code to estimate the Gaussian parameters ( and ), evaluate the probabilities for both the training data X from which you estimated the Gaussian parameters, and do so for the the cross-validation set Xval. Finally, it will use selectThreshold to find the best threshold *ϵ*. You should see a value epsilon of about 1.38e-18, and 117 anomalies found.

% Loads the second dataset. You should now have the variables X, Xval, yval in your environment

load('ex8data2.mat');

% Apply the same steps to the larger dataset

[mu, sigma2] = estimateGaussian(X);

% Training set

p = multivariateGaussian(X, mu, sigma2);

% Cross-validation set

pval = multivariateGaussian(Xval, mu, sigma2);

% Find the best threshold

[epsilon, F1] = selectThreshold(yval, pval);

fprintf('Best epsilon found using cross-validation: %e\n', epsilon);

fprintf('Best F1 on Cross Validation Set: %f\n', F1);

fprintf('# Outliers found: %d\n', sum(p < epsilon));

**2. Recommender Systems**

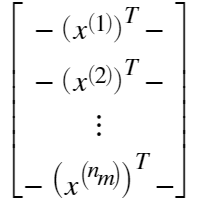
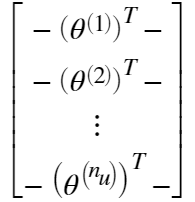
In this part of the exercise, you will implement the collaborative filtering learning algorithm and apply it to a dataset of movie ratings\*. This dataset consists of ratings on a scale of 1 to 5. The dataset has  users, and  movies. In the next parts of this exercise, you will implement the function cofiCostFunc.m that computes the collaborative fitlering objective function and gradient. After implementing the cost function and gradient, you will use fmincg.m to learn the parameters for collaborative filtering.

\*MovieLens 100k Dataset from GroupLens Research.

**2.1 Movie ratings dataset**

The code in this section will load the dataset ex8\_movies.mat, providing the variables Y and R in your MATLAB environment. The matrix Y (a num\_movies × num\_users matrix) stores the ratings  (from 1 to 5). The matrix R is an binary-valued indicator matrix, where R(i,j) = 1 if user j gave a rating to movie i, and R(i,j) = 0 otherwise. The objective of collaborative filtering is to predict movie ratings for the movies that users have not yet rated, that is, the entries with R(i,j) = 0. This will allow us to recommend the movies with the highest predicted ratings to the user.

To help you understand the matrix Y, the code below will compute the average movie rating for the first movie (Toy Story) and output the average rating to the screen. Throughout this part of the exercise, you will also be working with the matrices, X and Theta:

X = , Theta = .

% Load data

load('ex8\_movies.mat');

* Y is a 1682 x 943 matrix, containing ratings (1 - 5) of 1682 movies on 943 users
* R is a 1682 x 943 matrix, where R(i,j) = 1 if and only if user j gave a rating to movie i

% From the matrix, we can compute statistics like average rating.

fprintf('Average rating for movie 1 (Toy Story): %f / 5\n\n', mean(Y(1, R(1, :))));

% We can "visualize" the ratings matrix by plotting it with imagesc

imagesc(Y);

ylabel('Movies');

xlabel('Users');

The *i*-th row of X corresponds to the feature vector  for the *i*-th movie, and the *j*-th row of Theta corresponds to one parameter vector , for the *j*-th user. Both  and  are *n*-dimensional vectors. For the purposes of this exercise, you will use , and therefore,  and . Correspondingly, X is a  matrix and Theta is a  matrix.

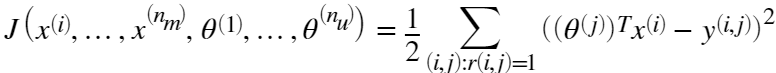
**2.2 Collaborative filtering learning algorithm**

Now, you will start implementing the collaborative filtering learning algorithm. You will start by implementing the cost function (without regularization). The collaborative filtering algorithm in the setting of movie recommendations considers a set of *n*-dimensional parameter vectors  and , where the model predicts the rating for movie *i* by user *j* as . Given a dataset that consists of a set of ratings produced by some users on some movies, you wish to learn the parameter vectors  that produce the best fit (minimizes the squared error).

You will complete the code in cofiCostFunc.m to compute the cost function and gradient for collaborative filtering. Note that the parameters to the function (i.e., the values that you are trying to learn) are X and Theta. In order to use an off-the-shelf minimizer such as fmincg, the cost function has been set up to unroll the parameters into a single vector params. You had previously used the same vector unrolling method in the neural networks programming exercise.

**2.2.1 Collaborative filtering cost function**

The collaborative filtering cost function (without regularization) is given by



You should now modify cofiCostFunc.m to return this cost in the variable J. Note that you should be accumulating the cost for user *j* and movie *i* only if . After you have completed the function, the code below will run your cost function. You should expect to see an output of 22.22.

**Implementation Note:** We strongly encourage you to use a vectorized implementation to compute J, since it will later by called many times by the optimization package fmincg. As usual, it might be easiest to first write a non-vectorized implementation (to make sure you have the right answer), and the modify it to become a vectorized implementation (checking that the vectorization steps don't change your algorithm's output). To come up with a vectorized implementation, the following tip might be helpful: You can use the R matrix to set selected entries to 0. For example, R.\*M will do an element-wise multiplication between M and R; since R only has elements with values either 0 or 1, this has the effect of setting the elements of M to 0 only when the corresponding value in R is 0. Hence, sum(sum(R.\*M)) is the sum of all the elements of M for which the corresponding element in R equals 1.

% Load pre-trained weights (X, Theta, num\_users, num\_movies, num\_features)

load('ex8\_movieParams.mat');

% Reduce the data set size so that this runs faster

num\_users = 4; num\_movies = 5; num\_features = 3;

X = X(1:num\_movies, 1:num\_features);

Theta = Theta(1:num\_users, 1:num\_features);

Y = Y(1:num\_movies, 1:num\_users);

R = R(1:num\_movies, 1:num\_users);

% Evaluate cost function

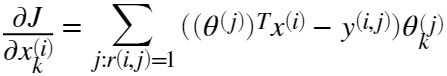
J = cofiCostFunc([X(:); Theta(:)], Y, R, num\_users, num\_movies,num\_features, 0);

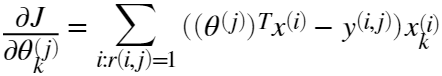
fprintf('Cost at loaded parameters: %f ',J);

*You should now submit your solutions. Enter* **submit***at the command prompt, then and enter or confirm your login and token when prompted.*

**2.2.2 Collaborative filtering gradient**

Now, you should implement the gradient (without regularization). Specically, you should complete the code in cofiCostFunc.m to return the variables X\_grad and Theta\_grad. Note that X\_grad should be a matrix of the same size as X and similarly, Theta\_grad is a matrix of the same size as Theta. The gradients of the cost function is given by:

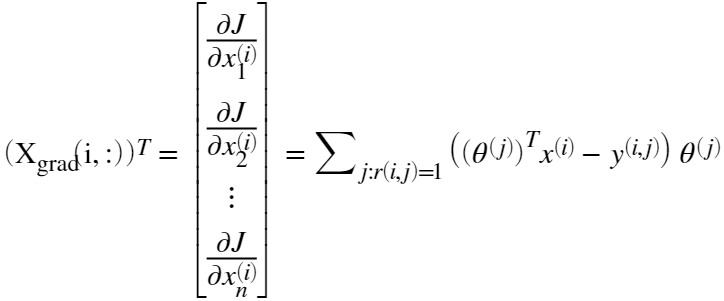




Note that the function returns the gradient for both sets of variables by unrolling them into a single vector. After you have completed the code to compute the gradients, the code below will run a gradient check (checkCostFunction) to numerically check the implementation of your gradients. (This is similar to the numerical check that you used in the neural networks exercise.) If your implementation is correct, you should find that the analytical and numerical gradients match up closely.

**Implementation Note**: You can get full credit for this assignment without using a vectorized implementation, but your code will run much more slowly (a small number of hours), and so we recommend that you try to vectorize your implementation. To get started, you can implement the gradient with a for loop over movies (for computing ) and a for loop over users (for computing ). When you first implement the gradient, you might start with an unvectorized version, by implementing another inner for loop that computes each element in the summation. After you have completed the gradient computation this way, you should try to vectorize your implementation (vectorize the inner for loops), so that you're left with only two for loops (one for looping over movies to compute  for each movie, and one for looping over users to compute  for each user).

To perform the vectorization, you might find this helpful: You should come up with a way to compute all the derivatives associated with , (i.e., the derivative terms associated with the feature vector ) at the same time. Let us define the derivatives for the feature vector of the *i*-th movie as:



To vectorize the above expression, you can start by indexing into Theta and Y to select only the elements of interest (that is, those with ). Intuitively, when you consider the features for the *i*-th movie, you only need to be concerned about the users who had given ratings to the movie, and this allows you to remove all the other users from Theta and Y.

Concretely, you can set idx = find(R(i,:)==1) to be a list of all the users that have rated movie *i*. This will allow you to create the temporary matrices Theta\_temp = Theta(idx,:) and Y\_temp = Y(i,idx) that index into Theta and Y to give you only the set of users which have rated the *i*-th movie. This will allow you to write the derivatives as:

.

(Note: The vectorized computation above returns a row-vector instead.) After you have vectorized the computations of the derivatives with respect to , you should use a similar method to vectorize the derivatives with respect to  as well.

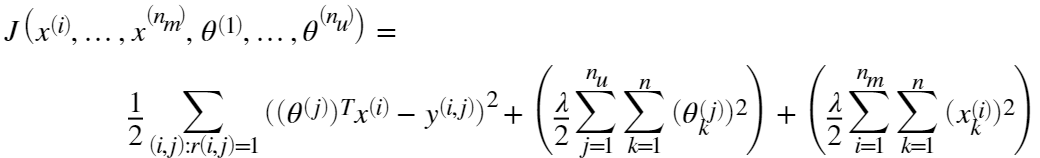
% Check gradients by running checkNNGradients

checkCostFunction;

*You should now submit your solutions. Enter* **submit***at the command prompt, then and enter or confirm your login and token when prompted.*

**2.2.3 Regularized cost function**

The cost function for collaborative filtering with regularization is given by



You should now add regularization to your original computations of the cost function, *J*. After you are done, the code below will run your regularized cost function, and you should expect to see a cost of about 31.34.

% Evaluate cost function

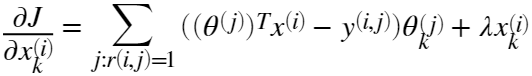
J = cofiCostFunc([X(:); Theta(:)], Y, R, num\_users, num\_movies, num\_features, 1.5);

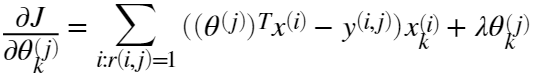
fprintf('Cost at loaded parameters (lambda = 1.5): %f',J);

*You should now submit your solutions. Enter* **submit***at the command prompt, then and enter or confirm your login and token when prompted.*

**2.2.4 Regularized gradient**

Now that you have implemented the regularized cost function, you should proceed to implement regularization for the gradient. You should add to your implementation in cofiCostFunc.m to return the regularized gradient by adding the contributions from the regularization terms. Note that the gradients for the regularized cost function is given by:





This means that you just need to add  to the X\_grad(i,:) variable described earlier, and add  to the Theta\_grad(j,:) variable described earlier. After you have completed the code to compute the gradients, the code below will run another gradient check (checkCostFunction) to numerically check the implementation of your gradients.

% Check gradients by running checkNNGradients

checkCostFunction(1.5);

*You should now submit your solutions. Enter* **submit***at the command prompt, then and enter or confirm your login and token when prompted.*

**2.3 Learning movie recommendations**

After you have finished implementing the collaborative ltering cost function and gradient, you can now start training your algorithm to make movie recommendations for yourself. In the code below, you can enter your own movie preferences, so that later when the algorithm runs, you can get your own movie recommendations! We have filled out some values according to our own preferences, but you should change this according to your own tastes. The list of all movies and their number in the dataset can be found listed in the file movie idx.txt.

% Load movvie list

movieList = loadMovieList();

% Initialize my ratings

my\_ratings = zeros(1682, 1);

% Check the file movie\_idx.txt for id of each movie in our dataset

% For example, Toy Story (1995) has ID 1, so to rate it "4", you can set

my\_ratings(1) = 4;

% Or suppose did not enjoy Silence of the Lambs (1991), you can set

my\_ratings(98) = 2;

% We have selected a few movies we liked / did not like and the ratings we gave are as follows:

my\_ratings(7) = 3;

my\_ratings(12)= 5;

my\_ratings(54) = 4;

my\_ratings(64)= 5;

my\_ratings(66)= 3;

my\_ratings(69) = 5;

my\_ratings(183) = 4;

my\_ratings(226) = 5;

my\_ratings(355)= 5;

fprintf('\n\nNew user ratings:\n');

for i = 1:length(my\_ratings)

if my\_ratings(i) > 0

fprintf('Rated %d for %s\n', my\_ratings(i), movieList{i});

end

end

**2.3.1 Recommendations**

After the additional ratings have been added to the dataset, the code below will proceed to train the collaborative filtering model. This will learn the parameters X and Theta.

% Load data

load('ex8\_movies.mat');

% Y is a 1682x943 matrix, containing ratings (1-5) of 1682 movies by 943 users

% R is a 1682x943 matrix, where R(i,j) = 1 if and only if user j gave a rating to movie i

% Add our own ratings to the data matrix

Y = [my\_ratings Y];

R = [(my\_ratings ~= 0) R];

% Normalize Ratings

[Ynorm, Ymean] = normalizeRatings(Y, R);

% Useful Values

num\_users = size(Y, 2);

num\_movies = size(Y, 1);

num\_features = 10;

% Set Initial Parameters (Theta, X)

X = randn(num\_movies, num\_features);

Theta = randn(num\_users, num\_features);

initial\_parameters = [X(:); Theta(:)];

% Set options for fmincg

options = optimset('GradObj','on','MaxIter',100);

% Set Regularization

lambda = 10;

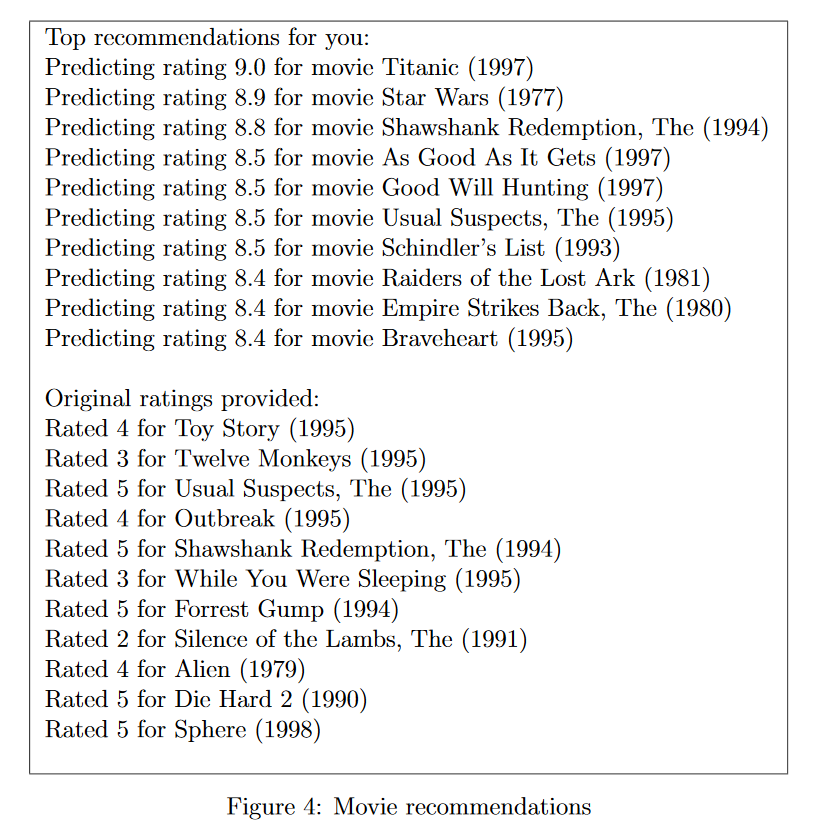
theta = fmincg(@(t)(cofiCostFunc(t, Ynorm, R, num\_users, num\_movies, num\_features,lambda)), initial\_parameters, options);

% Unfold the returned theta back into U and W

X = reshape(theta(1:num\_movies\*num\_features), num\_movies, num\_features);

Theta = reshape(theta(num\_movies\*num\_features+1:end), num\_users, num\_features);

To predict the rating of movie *i* for user*j*, you need to compute  The code below computes the ratings for all the movies and users and displays the movies that it recommends (Figure 4), according to ratings that were entered earlier in the script. Note that you might obtain a different set of the predictions due to different random initializations.



p = X \* Theta';

my\_predictions = p(:,1) + Ymean;

movieList = loadMovieList();

[r, ix] = sort(my\_predictions,'descend');

for i=1:10

j = ix(i);

if i == 1

fprintf('\nTop recommendations for you:\n');

end

fprintf('Predicting rating %.1f for movie %s\n', my\_predictions(j), movieList{j});

end

for i = 1:length(my\_ratings)

if i == 1

fprintf('\n\nOriginal ratings provided:\n');

end

if my\_ratings(i) > 0

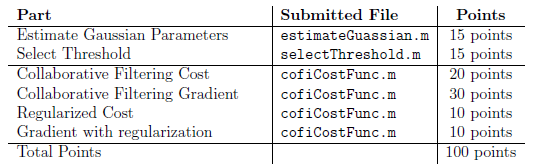
fprintf('Rated %d for %s\n', my\_ratings(i), movieList{i});

end

end

**Submission and Grading**

After completing various parts of the assignment, be sure to use the submit function system to submit your solutions to our servers. The following is a breakdown of how each part of this exercise is scored.



You are allowed to submit your solutions multiple times, and we will take only the highest score into consideration.