Programming Exercise 1: Linear Regression

Machine Learning

Introduction

In this exercise, you will implement linear regression and get to see it work on data. Before starting on this programming exercise, we strongly recommend watching the video lectures and completing the review questions for the associated topics.

To get started with the exercise, you will need to download the starter

code and unzip its contents to the directory where you wish to complete

the exercise. If needed, use the cd command in Octave to change to this directory before starting this exercise.

You can also find instructions for installing Octave on the “Octave Installation” page on the course website.

Files included in this exercise

ex1.m - Octave script that will help step you through the exercise ex1 multi.m - Octave script for the later parts of the exercise ex1data1.txt - Dataset for linear regression with one variable ex1data2.txt - Dataset for linear regression with multiple variables submit.m - Submission script that sends your solutions to our servers [⋆] warmUpExercise.m - Simple example function in Octave

[⋆] plotData.m - Function to display the dataset

[⋆] computeCost.m - Function to compute the cost of linear regression

[⋆] gradientDescent.m - Function to run gradient descent

[†] computeCostMulti.m - Cost function for multiple variables

[†] gradientDescentMulti.m - Gradient descent for multiple variables

[†] featureNormalize.m - Function to normalize features

[†] normalEqn.m - Function to compute the normal equations

⋆indicates files you will need to complete

† indicates extra credit exercises

Throughout the exercise, you will be using the scripts ex1.m and ex1 multi.m.

These scripts set up the dataset for the problems and make calls to functions that you will write. You do not need to modify either of them. You are only required to modify functions in other files, by following the instructions in this assignment.

For this programming exercise, you are only required to complete the first part of the exercise to implement linear regression with one variable. The second part of the exercise, which you may complete for extra credit, covers linear regression with multiple variables.

Where to get help

The exercises in this course use Octave,1 a high-level programming language well-suited for numerical computations. If you do not have Octave installed, please refer to the installation instructons at the “Octave Installation” page http://www.ml-class.org/course/resources/index?page=octave-install on the course website.

At the Octave command line, typing help followed by a function name

displays documentation for a built-in function. For example, help plot will bring up help information for plotting. Further documentation for Octave functions can be found at the Octave documentation pages.

We also strongly encourage using the online Q&A Forum to discuss discuss exercises with other students. However, do not look at any source code written by others or share your source code with others.

• Simple octave function

The first part of ex1.m gives you practice with Octave syntax and the home- work submission process. In the file warmUpExercise.m, you will find the outline of an Octave function. Modify it to return a 5 x 5 identity matrix by filling in the following code:

A = eye(5);

1Octave is a free alternative to MATLAB. For the programming exercises, you are free to use either Octave or MATLAB.

When you are finished, run ex1.m (assuming you are in the correct direc- tory, type “ex1” at the Octave prompt) and you should see output similar to the following:

ans =

Diagonal Matrix

1

0

0

0

0

0

1

0

0

0

0

0

1

0

0

0

0

0

1

0

0

0

0

0

1

Now ex1.m will pause until you press any key, and then will run the code for the next part of the assignment. If you wish to quit, typing ctrl-c will stop the program in the middle of its run.

• Submitting Solutions

After completing a part of the exercise, you can submit your solutions for grading by typing submit at the Octave command line. The submission script will prompt you for your username and password and ask you which files you want to submit. You can obtain a submission password from the website’s “Programming Exercises” page.

*You should now submit the warm up exercise.*

You are allowed to submit your solutions multiple times, and we will take only the highest score into consideration. To prevent rapid-fire guessing, the system enforces a minimum of 5 minutes between submissions.

All parts of this programming exercise are due Sunday, October 23rd

at 23:59:59 PDT.

• Linear regression with one variable

In this part of this exercise, you will implement linear regression with one variable to predict profits for a food truck. Suppose you are the CEO of a

restaurant franchise and are considering different cities for opening a new outlet. The chain already has trucks in various cities and you have data for profits and populations from the cities.

You would like to use this data to help you select which city to expand to next.

The file ex1data1.txt contains the dataset for our linear regression prob-

lem. The first column is the population of a city and the second column is the profit of a food truck in that city. A negative value for profit indicates a loss.

The ex1.m script has already been set up to load this data for you.

• Plotting the Data

Before starting on any task, it is often useful to understand the data by visualizing it. For this dataset, you can use a scatter plot to visualize the data, since it has only two properties to plot (profit and population). (Many other problems that you will encounter in real life are multi-dimensional and can’t be plotted on a 2-d plot.)

In ex1.m, the dataset is loaded from the data file into the variables *X*

and *y*:

data = csvread('ex1data1.txt'); X = data(:, 1); y = data(:, 2); m = length(y);

% read comma separated data

% number of training examples

Next, the script calls the plotData function to create a scatter plot of the data. Your job is to complete plotData.m to draw the plot; modify the file and fill in the following code:

plot(x, y, 'rx', 'MarkerSize', 10); % Plot the data ylabel('Profit in $10,000s'); % Set the y−axis label xlabel('Population of City in 10,000s'); % Set the x−axis label

Now, when you continue to run ex1.m, our end result should look like Figure 1, with the same red “x” markers and axis labels.

To learn more about the plot command, you can type help plot at the

Octave command prompt or to search online for plotting documentation. (To change the markers to red “x”, we used the option ‘rx’ together with the plot

command, i.e., plot(..,[your options here],.., ‘rx’); )

25

20

15

Profit in $10,000s

5

0

−5

4 6 8 10 12 14 16 18 20 22 24

Population of City in 10,000s

Figure 1: Scatter plot of training data

• Gradient Descent

In this part, you will fit the linear regression parameters *θ* to our dataset using gradient descent.

• Update Equations

The objective of linear regression is to minimize the cost function

*J* (*θ*) = 1

2*m*

*m*

*θ*

*i*=1

where the hypothesis *hθ* (*x*) is given by the linear model

*hθ* (*x*) = *θT x* = *θ*0 + *θ*1*x*1

Recall that the parameters of your model are the *θj* values. These are the values you will adjust to minimize cost *J* (*θ*). One way to do this is to use the batch gradient descent algorithm. In batch gradient descent, each iteration performs the update

*m*

1 (*i*)

(*i*)

(*i*)

*θj* := *θj* − *αm* X(*hθ* (*x*

*i*=1

) − *y*

)*xj* (simultaneously update *θj* for all *j*)*.*

With each step of gradient descent, your parameters *θj* come closer to the optimal values that will achieve the lowest cost *J* (*θ*).

Implementation Note: We store each example as a row in the the X

matrix in Octave. To take into account the intercept term (*θ*0), we add an additional first column to X and set it to all ones. This allows us to treat *θ*0 as simply another ‘feature’.

• Implementation

In ex1.m, we have already already set up the data for linear regression. In the following lines, we add another dimension to our data to accommodate the *θ*0 intercept term. We also initialize the initial parameters to 0 and the learning rate alpha to 0.01.

X = [ones(m, 1), data(:,1)]; % Add a column of ones to x theta = zeros(2, 1); % initialize fitting parameters

iterations = 1500;

alpha = 0.01;

• Computing the cost *J* (*θ*)

As you perform gradient descent to learn minimize the cost function *J* (*θ*), it is helpful to monitor the convergence by computing the cost. In this section, you will implement a function to calculate *J* (*θ*) so you can check the convergence of your gradient descent implementation.

Your next task is to complete the code in the file computeCost.m, which

is a function that computes *J* (*θ*). As you are doing this, remember that the variables *X* and *y* are not scalar values, but matrices whose rows represent the examples from the training set.

Once you have completed the function, the next step in ex1.m will run

computeCost once using *θ* initialized to zeros, and you will see the cost printed to the screen.

You should expect to see a cost of 32.07.

*You should now submit “compute cost” for linear regression with one variable.*

• Gradient descent

Next, you will implement gradient descent in the file gradientDescent.m. The loop structure has been written for you, and you only need to supply the updates to *θ* within each iteration.

As you program, make sure you understand what you are trying to opti- mize and what is being updated. Keep in mind that the cost *J* (*θ*) is parame- terized by the vector *θ*, not *X* and *y*. That is, we minimize the value of *J* (*θ*) by changing the values of the vector *θ*, not by changing *X* or *y*. Refer to the equations in this handout and to the video lectures if you are uncertain.

A good way to verify that gradient descent is working correctly is to look at the value of *J* (*θ*) and check that it is decreasing with each step. The starter code for gradientDescent.m calls computeCost on every iteration

and prints the cost. Assuming you have implemented gradient descent and

computeCost correctly, your value of *J* (*θ*) should never increase, and should converge to a steady value by the end of the algorithm.

After you are finished, ex1.m will use your final parameters to plot the linear fit. The result should look something like Figure 2:

Your final values for *θ* will also be used to make predictions on profits in areas of 35,000 and 70,000 people. Note the way that the following lines in ex1.m uses matrix multiplication, rather than explicit summation or loop- ing, to calculate the predictions. This is an example of code vectorization in Octave.

*You should now submit gradient descent for linear regression with one variable.*

predict1 = [1, 3.5] \* theta;

predict2 = [1, 7] \* theta;

• Debugging

Here are some things to keep in mind as you implement gradient descent:

❼ Octave array indices start from one, not zero. If you’re storing *θ*0 and

*θ*1 in a vector called theta, the values will be theta(1) and theta(2).

Training data Linear regression

20

15

Profit in $10,000s

5

0

−5

4 6 8 10 12 14 16 18 20 22 24

Population of City in 10,000s

Figure 2: Training data with linear regression fit

❼ If you are seeing many errors at runtime, inspect your matrix operations to make sure that you’re adding and multiplying matrices of compat- ible dimensions. Printing the dimensions of variables with the size command will help you debug.

❼ By default, Octave interprets math operators to be matrix operators. This is a common source of size incompatibility errors. If you don’t want matrix multiplication, you need to add the “dot” notation to specify this to Octave. For example, A\*B does a matrix multiply, while A.\*B does an element-wise multiplication.

• Visualizing *J* (*θ*)

To understand the cost function *J* (*θ*) better, you will now plot the cost over a 2-dimensional grid of *θ*0 and *θ*1 values. You will not need to code anything new for this part, but you should understand how the code you have written already is creating these images.

In the next step of ex1.m, there is code set up to calculate *J* (*θ*) over a

grid of values using the computeCost function that you wrote.

% initialize J vals to a matrix of 0's

J vals = zeros(length(theta0 vals), length(theta1 vals));

% Fill out J vals

for i = 1:length(theta0 vals)

for j = 1:length(theta1 vals)

t = [theta0 vals(i); theta1 vals(j)]; J vals(i,j) = computeCost(x, y, t);

end

end

After these lines are executed, you will have a 2-D array of *J* (*θ*) values. The script ex1.m will then use these values to produce surface and contour plots of *J* (*θ*) using the surf and contour commands. The plots should look something like Figure 3:

4

3.5

800

700

600

500

400

300

3

2.5

2



1

200 1

100

0

4

3

2

1

0

 −1

1

5

0

−5

−10 

0

0.5

0

10

−0.5

−1

−10 −8 −6 −4 −2 0 2 4 6 8 10



0

• Surface

•   
Contour, showing minimum

Figure 3: Cost function *J* (*θ*)

The purpose of these graphs is to show you that how *J* (*θ*) varies with changes in *θ*0 and *θ*1. The cost function *J* (*θ*) is bowl-shaped and has a global mininum. (This is easier to see in the contour plot than in the 3D surface plot). This minimum is the optimal point for *θ*0 and *θ*1, and each step of gradient descent moves closer to this point.

Extra Credit Exercises (optional)

If you have successfully completed the material above, congratulations! You now understand linear regression and should able to start using it on your own datasets.

For the rest of this programming exercise, we have included the following optional extra credit exercises. These exercises will help you gain a deeper understanding of the material, and if you are able to do so, we encourage you to complete them as well.

• Linear regression with multiple variables

In this part, you will implement linear regression with multiple variables to predict the prices of houses. Suppose you are selling your house and you want to know what a good market price would be. One way to do this is to first collect information on recent houses sold and make a model of housing prices.

The file ex1data2.txt contains a training set of housing prices in Port-

land, Oregon. The first column is the size of the house (in square feet), the second column is the number of bedrooms, and the third column is the price of the house.

The ex1 multi.m script has been set up to help you step through this

exercise.

• Feature Normalization

The ex1 multi.m script will start by loading and displaying some values from this dataset. By looking at the values, note that house sizes are about 1000 times the number of bedrooms. When features differ by orders of mag- nitude, first performing feature scaling can make gradient descent converge much more quickly.

Your task here is to complete the code in featureNormalize.m to

❼ Subtract the mean value of each feature from the dataset.

❼ After subtracting the mean, additionally scale (divide) the feature values by the *inverse* of their respective “standard deviations.”

The standard deviation is a way of measuring how much variation there is in the range of values of a particular feature (most data points will lie

within ±2 standard deviations of the mean); this is an alternative to taking the range of values (max-min). In Octave, you can use the “std” function to compute the standard deviation. For example, inside featureNormalize.m,

the quantity X(:,1) contains all the values of *x*1 (house sizes) in the training set, so std(X(:,1)) computes the standard deviation of the house sizes. At the time that featureNormalize.m is called, the extra column of 1’s corresponding to *x*0 = 1 has not yet been added to X (see ex1 multi.m for details).

You will do this for all the features and your code should work with datasets of all sizes (any number of features / examples). Note that each column of the matrix X corresponds to one feature.

*You should now submit feature normalization.*

Implementation Note: When normalizing the features, it is important

to store the values used for normalization - the *mean value* and the *stan- dard deviation* used for the computations. After learning the parameters from the model, we often want to predict the prices of houses we have not

seen before. Given a new x value (living room area and number of bed-

rooms), we must first normalize x using the mean and standard deviation that we had previously computed from the training set.

• Gradient Descent

Previously, you implemented gradient descent on a univariate regression problem. The only difference now is that there is one more feature in the matrix X. The hypothesis function and the batch gradient descent update rule remain unchanged.

You should complete the code in computeCostMulti.m and gradientDescentMulti.m

to implement the cost function and gradient descent for linear regression with multiple variables. If your code in the previous part (single variable) already supports multiple variables, you can use it here too.

Make sure your code supports any number of features and is well-vectorized. You can use ‘size(X, 2)’ to find out how many features are present in the

dataset.

*You should now submit compute cost and gradient descent for linear re- gression with multiple variables.*

Implementation Note: In the multivariate case, the cost function can

also be written in the following vectorized form:

*J* (*θ*) = (*Xθ* − *~y*)*T* (*Xθ* − *~y*)

1

2*m*

where



— (*x*(1))*T* —





*y*(1)



*X* = 

— (*x* )

(2) *T*

—

*y*





...







— (*x*(*m*))*T* — 



*~y* =









(2)

...



*.*





*y*(*m*)



The vectorized version is efficient when you’re working with numerical

computing tools like Octave. If you are an expert with matrix operations, you can prove to yourself that the two forms are equivalent.

3.2.1 Optional (ungraded) exercise: Selecting learning rates

In this part of the exercise, you will get to try out different learning rates for the dataset and find a learning rate that converges quickly. You can change the learning rate by modifying ex1 multi.m and changing the part of the code that sets the learning rate.

The next phase in ex1 multi.m will call your gradientDescent.m func- tion and run gradient descent for about 50 iterations at the chosen learning rate. The function should also return the history of *J* (*θ*) values in a vector

J. After the last iteration, the ex1 multi.m script plots the J values against the number of the iterations.

If you picked a learning rate within a good range, your plot look similar Figure 4. If your graph looks very different, especially if your value of *J* (*θ*) increases or even blows up, adjust your learning rate and try again. We rec- ommend trying values of the learning rate *α* on a log-scale, at multiplicative steps of about 3 times the previous value (i.e., 0.3, 0.1, 0.03, 0.01 and so on). You may also want to adjust the number of iterations you are running if that will help you see the overall trend in the curve.

Figure 4: Convergence of gradient descent with an appropriate learning rate

Implementation Note: If your learning rate is too large, *J* (*θ*) can di- verge and ‘blow up’, resulting in values which are too large for computer calculations. In these situations, Octave will tend to return NaNs. NaN

stands for ‘not a number’ and is often caused by undefined operations that involve −∞ and +∞.

Octave Tip: To compare how different learning learning rates affect convergence, it’s helpful to plot J for several learning rates on the same figure. In Octave, this can be done by performing gradient descent multi- ple times with a ‘hold on’ command between plots. Concretely, if you’ve tried three different values of alpha (you should probably try more values than this) and stored the costs in J1, J2 and J3, you can use the following commands to plot them on the same figure:

plot(1:50, J1(1:50), ‘b’);

hold on;

plot(1:50, J2(1:50), ‘r’);

plot(1:50, J3(1:50), ‘k’);

The final arguments ‘b’, ‘r’, and ‘k’ specify different colors for the plots.

Notice the changes in the convergence curves as the learning rate changes. With a small learning rate, you should find that gradient descent takes a very long time to converge to the optimal value. Conversely, with a large learning rate, gradient descent might not converge or might even diverge!

Using the best learning rate that you found, run the ex1 multi.m script

to run gradient descent until convergence to find the final values of *θ*. Next, use this value of *θ* to predict the price of a house with 1650 square feet and 3 bedrooms. You will use value later to check your implementation of the normal equations. Don’t forget to normalize your features when you make this prediction!

*You do not need to submit any solutions for these optional (ungraded) exercises.*

3.3 Normal Equations

In the lecture videos, you learned that the closed-form solution to linear regression is

1

*θ* = *XT X*

*XT ~y.*

Using this formula does not require any feature scaling, and you will get an exact solution in one calculation: there is no “loop until convergence” like in gradient descent.

Complete the code in normalEqn.m to use the formula above to calcu-

late *θ*. Remember that while you don’t need to scale your features, we still need to add a columns of 1’s to the X matrix to have an intercept term (*θ*0).

*You should now submit the normal equations function.*

*Optional (ungraded) exercise:* Now, once you have found *θ* using this method, use it to make a price prediction for a 1650-square-foot house with 3 bedrooms. You should find that gives the same predicted price as the value you obtained using the model fit with gradient descent (in Section 3.2.1).

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.

Part

Submitted File

Points

Warm up exercise

warmUpExercise.m

10 points

Compute cost for one variable Gradient descent for one variable

computeCost.m

gradientDescent.m

40 points

50 points

Total Points

100 points

Extra Credit Exercises (optional)

Feature normalization Compute cost for multiple variables

Gradient descent for multiple variables

featureNormalize.m

computeCostMulti.m gradientDescentMulti.m

10 points

15 points

15 points

Normal Equations

normalEqn.m

10 points

You are allowed to submit your solutions multiple times, and we will take the highest score into consideration. To prevent rapid-fire guessing, the system enforces a minimum of 5 minutes between submissions.

All parts of this programming exercise are due Sunday, October 23rd

at 23:59:59 PDT.