Multivariate Linear Regression

Objectives

- To investigate multivariate linear regression using gradient descent and the normal equations.
- To examine the relationship between the cost function $J(\theta)$ the convergence of gradient descent, and the learning rate α.

Data

- Download ex2Data.zip, and extract the files from the zip file.
- This is a training set of housing prices in Portland, Oregon, where the outputs y(i) are the prices and the inputs x(i) are the living area and the number of bedrooms.
- There are m = 47 training examples.

Procedure 2.0

1. Plot the raw data: (a) housing prices with respect to living area; (b) housing prices with respect to the number of bedrooms before pre-processing.

2. Plot the preprocessed data: (a) housing prices with respect to living area; (b) housing prices with respect to the number of bedrooms after pre-processing.

Preprocessing your data

- Load the data for the training examples into your program and add the $x_0 = 1$ intercept term into your x matrix.
- Recall that the command in Matlab/Octave for adding a column of ones is

```
x = [ones(m, 1), x];
```

Preprocessing your data

- Take a look at the values of the inputs x(i) and note that the living areas are about 1000 times the number of bedrooms.
- This difference means that preprocessing the inputs will significantly increase gradient descent's efficiency.

Preprocessing your data

- In your program, scale both types of inputs by their standard deviations and set their means to zero.
- In Matlab/Octave, this can be executed with

```
sigma = std(x);
mu = mean(x);
x(:,2) = (x(:,2) - mu(2))./ sigma(2);
x(:,3) = (x(:,3) - mu(3))./ sigma(3);
```

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 is still

$$h_\theta(x) = \theta^T x = \sum_{i=0}^{n} \theta_i x_i,$$

and the batch gradient descent update rule is

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)} \qquad \text{(for all } j\text{)}$$

Initialization

- Once again, initialize your parameters to

$$\theta = \vec{0}$$

Selecting a learning rate using *J(θ)*

- The 1st goal is to pick a good learning rate in the range of

$$0.001 \leq \alpha \leq 10$$

- You will do this by making an initial selection, running gradient descent and observing the cost function, and adjusting the learning rate accordingly.
- Recall that the cost function is defined as

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Vectorized Cost function

- The cost function can also be written in the following vectorized form,

$$J(\theta) = \frac{1}{2m} (X\theta - \vec{y})^T (X\theta - \vec{y})$$

where:

$$\vec{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(m)} \end{bmatrix} \quad X = \begin{bmatrix} —(x^{(1)})^T— \\ —(x^{(2)})^T— \\ \vdots \\ —(x^{(m)})^T— \end{bmatrix}$$

Vectorized Cost function

- The vectorized version is useful and efficient when you're working with numerical computing tools like Matlab/Octave.
- If you are familiar with matrices, you can prove to yourself that the two forms are equivalent.

While in the previous exercise you calculated $J(\theta)$ over a grid of θ_0 and θ_1 values, you will now calculate $J(\theta)$ using the θ of the current stage of gradient descent. After stepping through many stages, you will see how $J(\theta)$ changes as the iterations advance.

Now, run gradient descent for about 50 iterations at your initial learning rate. In each iteration, calculate $J(\theta)$ and store the result in a vector J. After the last iteration, plot the J values against the number of the iteration. In Matlab/Octave, the steps would look something like this:

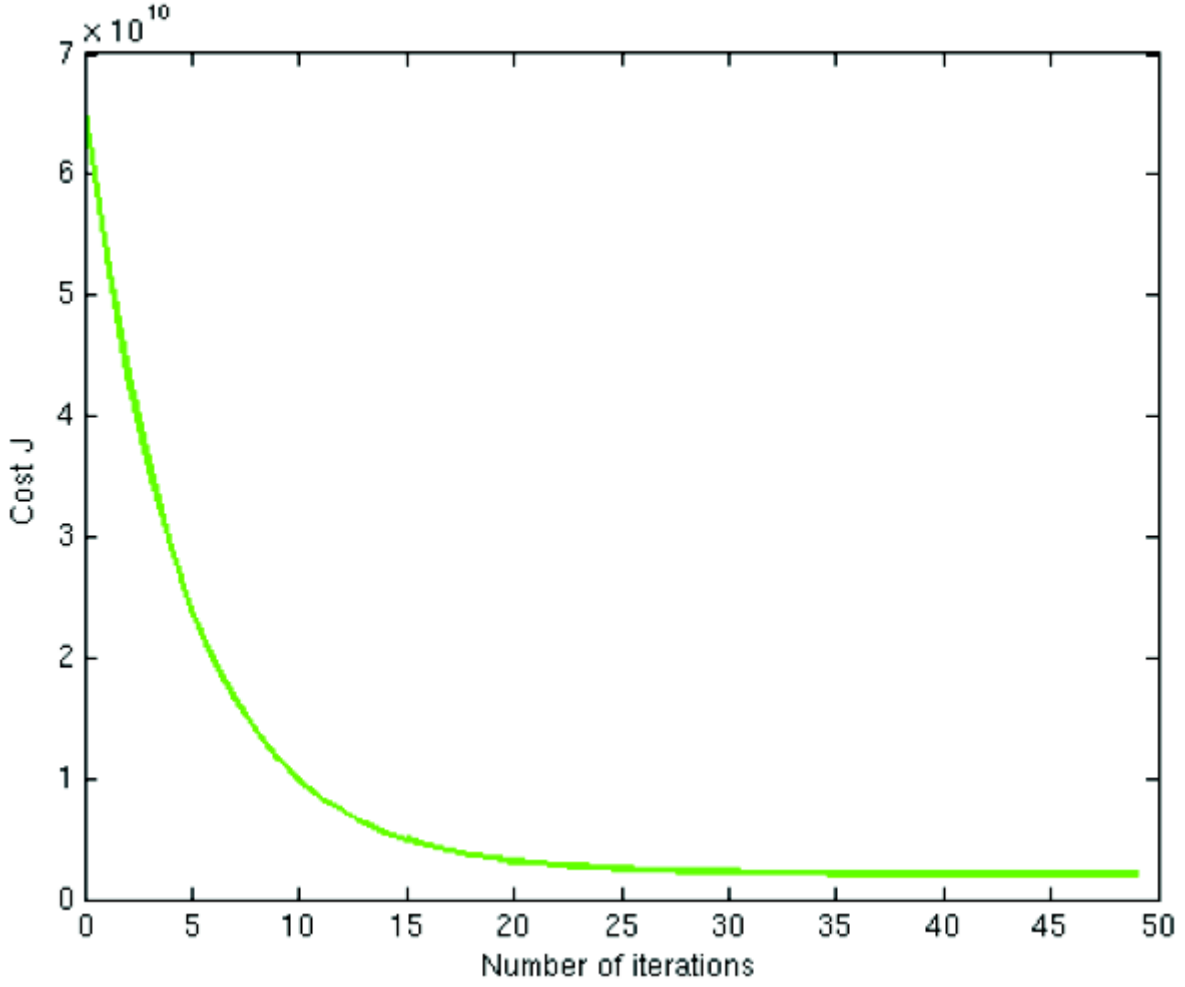
```
theta = zeros(size(x(1,:)))'; % initialize fitting parameters
alpha = %% Your initial learning rate %%
J = zeros(50, 1);

for num_iterations = 1:50
    J(num_iterations) = %% Calculate your cost function here %%
    theta = %% Result of gradient descent update %%
end

% now plot J
% technically, the first J starts at the zero-eth iteration
% but Matlab/Octave doesn't have a zero index
figure;
plot(0:49, J(1:50), '-')
xlabel('Number of iterations')
ylabel('Cost J')
```

Output 1

- If you picked a learning rate within a good range, your plot should appear like the figure below.



Tuning

- If your graph looks very different, especially if your value of $J(\theta)$ increases or even blows up, adjust your learning rate and try again.
- We recommend testing alphas at a rate of of 3 times the next smallest value (i.e. 0.01, 0.03, 0.1, 0.3 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.

Comparison Plot

- To compare how different learning learning rates affect convergence, it's helpful to plot J for several learning rates on the same graph.
- In Matlab/Octave, this can be done by performing gradient descent multiple 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(0:49, J1(1:50), 'b-');
hold on;
plot(0:49, J2(1:50), 'r-');
plot(0:49, J3(1:50), 'k-');
```

Plotting

- The final arguments `b', `r', and 'k' specify different plot styles for the plots.
- Type

```
help plot
```

  at the Matlab/Octave command line for more information on plot styles.

```
>> help plot
'plot' is a function from the file /usr/share/octave/4.0.0/m/plot/draw/plot.m

 -- Function File:  plot (Y)
 -- Function File:  plot (X, Y)
 -- Function File:  plot (X, Y, FMT)
 -- Function File:  plot (..., PROPERTY, VALUE, ...)
 -- Function File:  plot (X1, Y1, ..., XN, YN)
 -- Function File:  plot (HAX, ...)
 -- Function File: H = plot (...)
     Produce 2-D plots.

     Many different combinations of arguments are possible.  The
     simplest form is

          plot (Y)

     where the argument is taken as the set of Y coordinates and the X
     coordinates are taken to be the range `1:numel (Y)'.

     If more than one argument is given, they are interpreted as

          plot (Y, PROPERTY, VALUE, ...)
```

Procedure 2.1

- Observe the changes in the cost function happens as the learning rate changes. What happens when the learning rate is too small? too large?
- What is the best learning rate that you have found?

Procedure 2.2

- Using the best learning rate that you found, run gradient descent until convergence to find

  1. The final values of θ = _____

  2. The predicted price of a house with 1650 square feet and 3 bedrooms.

  (Don't forget to scale your features when you make this prediction!)

  House price = _____

Normal Equations

- The closed form solution to a least squares fit is

$$\theta = \left(X^T X\right)^{-1} X^T \vec{y}.$$

- 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.

Procedure 2.3

- In your program, use the closed form solution to a least squares fit formula to calculate θ.
- Remember that while you don't need to scale your features, you still need to add an intercept term.
- Once you have found from this method, use it to make a price prediction for a 1650 squarefoot house with 3 bedrooms.
- Did you get the same price that you found through gradient descent?

Procedure 2.4

- Provide your own data with at least 50 elements and at least 2 input features. Then, apply the procedures 2.1-2.3 performed in this experiment to your own data.

End

Reference:

- Andrew Ng. Stanford University, CS 229 Machine LearningCourse Materials. http://cs229.stanford.edu/materials.html