Module Big Data & Small Data
Apply supervised machine learning
Linear regression

Description of assignment

In this assignment you have to implement linear regression using gradient descent. The provided dataset (auto-mpg.mat) contains data of a set of cars with which we want to build a model for predicting the mileage per gallon. You have to do the following:

(1) Implement a MATLAB function: 'mvgd' (multi-variate gradient descent), i.e.,

```
[theta,cost_h] = mvgd(X,y,theta_0,alpha,interations)
```

The function 'mvgd' must accept:

- 'X' : the feature matrix,
- 'y' : the outcomes vector,
- 'theta_0' : the initial theta vector,
- 'alpha' : the learning rate,
- 'iterations' : number of iterations,

as its *input* arguments.

And, the function must return two *outputs*, namely,

- 'theta' : the (computed) final theta vector,
- 'cost_h' : a vector containing the values of the cost function from each iteration step.

(2) Implement a feature normalization function, so-called 'normalizeFeatures' , i.e.

```
[norm_X,mu,sigma] = normalizeFeatures(X)
```

The function 'normalizeFeatures' must accept

- 'X' : the feature matrix,

and return the following outputs:

- 'norm_X' : the normalized feature matrix,
- 'mu' and 'sigma' : the parameters used for feature normalization.

(3) Analyze the effects of feature normalization

Without normalization, the given combination of learning rate (alpha = 0.3) and iterations (iterations =1000) does not result in a convergence of the learning process.

(i) Explain why this phenomenon happens.

(ii) Find values of ‘alpha’ and ‘iterations’ that lead to convergence.
(iii) Compare these findings to the case of learning with *normalized* features.

(4) ‘Direct’ VS ‘iterative’ solutions for Linear regression

Note that, for the provided dataset, it is also possible to solve this linear regression problem by using the *normal equation* directly.

With the Matlab commands `tic` and `toc` you can start and stop a stopwatch to measure the time it takes to execute a set of commands.

Compare the results, model parameters and execution time, that you have acquired through the *gradient descent* algorithm with the results obtained from the *normal equation*. Discuss and summarize your observations.

Required files

auto-mpg.mat

run_mvlr.m