EEL709: Assignment 1

Weightage: 10%

February 10, 2015

1 Objective

The objective here is to implement the concepts of regression learnt in class via polynomial curve fitting. To recap, polynomial curve fitting is an example of regression. In regression, the objective is to learn a function that maps an input variable x to a continuous target variable t.

For the first part of this assignment, we provide a personalised input file that contains data of the form:

$$\begin{array}{l} x_1, t_1 \\ x_2, t_2 \\ . \\ . \\ . \\ x_{100}, t_{100} \end{array}$$

The relation between x and t is of the form

$$t = w_0 + w_1 x + ... + w_M x^M + \epsilon$$

where the noise ϵ is drawn from a normal distribution with mean 0 and unknown (but fixed, for a given file) variance. M is also unknown. You should download your file from `http://web.iitd.ac.in/~sumeet/A1/<EntryID>.csv` (for example, `http://web.iitd.ac.in/~sumeet/A1/2010EE50541.csv`). The end goal is to identify the underlying polynomial (both the degree and the coefficients), as well as to obtain an estimate of the noise variance.

The tasks to be accomplished are:

- To begin with, use only the first 20 data points in your file.
- You may use standard functions/libraries in any programming language of your choice: MATLAB, Java, Python etc.
- Solve the curve fitting regression problem using error function minimisation. You can define your own error function other than sum-of-squares error (note that the error function need not be convex). Try different error formulations and report the results. Also try and use a validation approach to characterise the goodness of fit for polynomials of different order. Can you distinguish overfitting, underfitting, and the best fit? In addition to this, obtain an estimate for the noise variance.
- Introduce regularisation and observe the changes. For quadratic regularisation, can you obtain an estimate of the optimal value for the regularisation parameter λ? What is your corresponding best guess for the underlying polynomial? And the noise variance? (For extra credit: compare these results with what you get by using lasso regularisation.)

- Now repeat all of the above using the full data set of 100 points. How are your results affected by adding more data? Comment on the differences.
- At the end: what is your final estimate of the true underlying polynomial? Why?

For the second part of the assignment you have to:

- Obtain a publicly available dataset for multi-variable linear regression (i.e., multiple input features, but a single target output). Possible sources are `https://archive.ics.uci.edu/ml/datasets.html`, `http://www.kaggle.com`, `http://people.sc.fsu.edu/~jburkardt/datasets/regression/regression.html`.
- Understand the importance of the different input features.
- Come up with the best linear regression model you can, and interpret its performance.
- Visualise the results and variations on omitting some of the features.

2 Evaluation

- You are required to give a demonstration of regression, the coefficients you've obtained and how you've done so. [You might be asked to reproduce some of the results.]
- Visualise the data and results in meaningful ways. [Gnuplot is a good tool for this.]

2.1 What we'll be looking for

- For the first part, how close did you get to the actual answer? (Which we know, and will tell you during your viva!) If there is a big discrepancy, why might this have happened?
- How well you understood what you are demonstrating and the concepts involved.
- How extensively you have played around with various parameters, and your analysis of the variations in the consequent results.
- The insights that you have gained from your experiments.