CS1762: Experiment 1

Maximum Marks: 10

Submission deadline: 16 August, 2020

1. Write a program to implement the concepts of regression learnt in class via polynomial curve fitting.

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. we provide a personalized input file that contains data of the form: $\{(x_1, t_1), (x_2, t_2)........(x_{100}, t_{100})\}$. The relation between x and t is of the form

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

Where $\in$ the noise is drawn from a normal distribution with mean 0 and unknown (but fixed, for a given file) variance. M is also unknown. 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 minimization. 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 characterize the goodness of fit for polynomials of different order. Can you distinguish over-fitting, under-fitting, and the best fit? In addition to this, obtain an estimate for the noise variance.

- Introduce regularization and observe the changes. For quadratic regularization, can you obtain an estimate of the optimal value for the regularization parameter λ? What is your corresponding best guess for the underlying polynomial? And the noise variance.

- 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?