ELL784/EEL709: Assignment 1

Maximum Marks: 6+2

Submission deadline: **15 February, 23:59**

1 Part 1 (3 marks)

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:

$$\begin{aligned} &x_1, t_1 \\ &x_2, t_2 \\ &. \\ &. \\ &. \\ &x_{100}, t_{100} \end{aligned}$$

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>/Gaussian_noise` (for example, `http://web.iitd.ac.in/~sumeet/A1/2008BB50007/Gaussian_noise`). 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?

- 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?

1.1 Part 1EC (extra credit, 2 marks)

You are provided with a second personalised data set, available at `http://web.iitd.ac.in/~sumeet/A1/<EntryID>/NonGaussian_noise`. It is generated from the same type of polynomial, except that the noise is now non-Gaussian. Can you repeat the analysis for this data set, focusing in particular on characterising the noise – can you figure out what kind of noise it is? Please justify your answer using appropriate analysis.

1.2 Evaluation criteria

- 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.]
- 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.

2 Part 2 (3 marks)

The Lockheed Martin Falcon is a single-engine multirole fighter aircraft developed for the United States Air Force. To control a fighter aircraft, an interesting exercise would be the prediction and control of the elevators, which are flight control surfaces, usually at the rear of an aircraft, adjusting the aircraft's pitch, and therefore the angle of attack and the lift of the wing.

We have provided a training data set (URL below) involving 18 different operational parameters (Roll rate, climb rate etc.) for the Falcon aircraft, and the target is to predict the elevator variable at any instant using a multivariate linear regression model.

You are allowed to use only multivariate linear regression models for this task. Cross-validation, hyperparameter tuning and regularization are encouraged to produce better results.

2.1 Evaluation criteria

Evaluation on this task will be based on conceptual clarity, interpretation of results and reproducibility. Your understanding of the algorithm, and the extent of validation and tuning will also be evaluated.

Additionally, you are also provided with a test set (without labels). You have to predict the label for this set using the model you have trained from the training set, and submit your results for a real-time leaderboard at `http://dubeya.com/ell784/`, where you will be ranked on your performance on the test set.

2.2 Data

- `http://web.iitd.ac.in/~sumeet/A1/train.csv`, A CSV file with one row per instance, and 18 features followed by the label. Total number of instances is 3,369.
- `http://web.iitd.ac.in/~sumeet/A1/test.csv`, A CSV file with one row per instance, and 18 features. There is no label at the end. Total number of instances is 706.

For submission to the leaderboard, you will be emailed a secret key which will be required at submission. Your IIT Kerberos ID is your user ID for the leaderboard. The format for submission is 706 comma-separated labels, without spaces or line-breaks. For example, if the test data had 5 instances, a valid submission would be 1.1,4.5,9.003,2,1.442.

The leaderboard will be accepting submissions till the assignment deadline.

3 Evaluation

- You should prepare a report, compiling all your results and your interpretation of them, along with your overall conclusions. In particular, you should attempt to answer all of the questions posed above. Any graphs or other visualisations should also be included therein. If you wish, you may also include code or other materials which are relevant, though this is not required. The submission link is `http://web.iitd.ac.in/~sumeet/submit.html`; put everything into a single zip file or tarball, and name it as per the instructions given there. The submission deadline is **February 15th, 23:59**. Any late submissions will be penalised.
- The schedule for demos/vivas will be announced by your respective TAs, in advance. If for any reason you cannot attend in your scheduled slot, you must arrange for an alternative slot with your TA well in advance. Last-minute requests for rescheduling will normally not be accepted.