Machine Learning (BITS F464) - Assignment 1

Polynomial Regression

Maximum Marks: 40

Submission Deadline: 0900Hrs 26/06/2018

Generate Dataset

Let f be a function from [0,1] to [-1,1] defined by f(x) = sin 2∏x. Suppose € follows normal distribution with mean 0 and variance ~~1~~ 0.1 i.e., N(0,~~1~~ 0.1). Randomly draw K points from [0,1]. For each of these random values, x, generate L target responses as $f(x) + €_1$, $f(x) + €_2$,. . . , $f(x) + €_L$ where $€_1$,$€_2$, . . . , $€_L$ are drawn from N(0,1). For each random value, L examples can be generated resulting total of KL instances.

Problem 1

Part A: Generate training dataset of size 20 as mentioned above.

Let $y(x,w) = w_0 + w_1 x^1 + w_2 x^2 + . . . + w_D x^D$ be a polynomial degree D.

Using least squares approach, fit the polynomial of degree 0, 1, 2, . . ., 10 for the training data.

Appreciate and document over fitting in the models developed above.

Part B: Generate training dataset of suitable size (as maximum as your laptop can handle) as mentioned above.

Let $y(x,w) = w_0 + w_1 x^1 + w_2 x^2 + . . . + w_D x^D$ be a polynomial degree D.

Using least squares approach, fit the polynomial of degree 0, 1, 2, . . ., 10 for the training data.

Will there be overfitting in the models developed above and write your comments in the report.

Problem 2

Using the regularization as the technique to combat overfitting, apply quadratic regularizer to the training data that is generated in Problem 1 – Part A.

Find the optimal λ for each of the polynomials of degree 0, 1, 2, . . . , 10.

Provide supporting graphs for selecting the optimal λ .

Problem 3

Using Bayesian curve fitting as discussed in class, fit polynomials of degree 0, 1, 2, . . . , 10. The training data generated for Problem 1 – Part A should be used to build these models.

Problem 4

Put up a comparative study of performance of models developed in the above three problems on the training data (generated for Problem 1 – Part A).

Generate testing data (as mentioned in ***Generate Dataset*** section) with number of instances equal to 1/3rd of the training instances.

Put up a comparative study of performance of models developed in the above three problems on the testing data.

Analyse your results and discuss the reasons for differences in performance of the models.

Packages Allowed: R / Python

Report:

- Team Members
- Methodology
- Supporting Documentation – Graphs, Tables etc.
- Analysis
- Results & Discussion

Evaluation:

- Viva
- Final results
- Understanding of results.
- Ability to reason the derived results.
- Final report and demo.

Submission should be through **CMS** only.