CS F320 Foundations of Data Science

Assignment-1

Submission Time & Date: 13:00hrs on 20th Oct 2023 ~~xxxxxxxxx~~ 7th Oct 23

Instructions

- This assignment is a coding project and is expected to be done in groups. Each group can contain at most three members. Make sure that all members in the group are registered to this course.
- This assignment is expected to be done in Python using standard libraries like NumPy, Pandas and Matplotlib. You are not supposed to use libraries like scikit-learn for the regression models. Jupyter Notebook/Google Colab can be used. Refrain from directly copying codes/snippets from other groups or the internet as all codes will be put through a plagiarism check.
- All deliverable items (ex. .py files, .ipynb files, reports, images) should be put together in a single .zip file. Rename this file as A1_<id-of-first-member>_ <id-of-second- member>_ <id-of-third-member> before submission.
- Submit the zip file on CMS on or before the aforementioned deadline. Please note that this is a hard deadline and no extensions/exemptions will be given. The demos for this assignment will be held later which shall be conveyed to you. All group members are expected to be present during the demo.

Assignment 1-A : Regression without regularization

Problem Statement

- The given dataset is a synthetic dataset consisting of 1,000 data points, each having one feature variable and one continuous target variable.

https://drive.google.com/file/d/14GTYF60OdJzGK8IopRGqpx1ZPncrjCNX/view?usp=sharing

Independent Variable : X , Dependent Variable : Y

- **Task 1: Data Preprocessing**

    a. Load the shared dataset into a pandas DataFrame.

    b. Normalize the feature variable by utilizing the formula: X' = (X - μ) / σ where μ represents the mean of feature value, and σ represents the standard deviation of feature values.

    c. Shuffle the dataset and split the dataset into training and testing sets (80% for training and 20% for testing).

- **Task 2: Polynomial Regression**

    a. Build polynomial regression models (with degrees varying from 1 to 9) to predict the target variable based on the input feature variable. Determine the degree of the polynomial which best fits the given data.

    b. Apply batch gradient descent to train the models.

    c. Train each model for 500 iterations.

Note : The value of the learning rate can be varied and the best result should be documented.

- **Task 3: Graph Plotting**

    a. Plot 1 - Final Training and Testing Errors v/s degree of polynomial

    b. Training Error and Testing Error v/s Epochs for all degree of polynomials in [1, 9]

    c. The best polynomial fitted curve on the data points. (Choose a subset of points for better visualization)

- **Task 4: Comparative Analysis**

    a. Perform the comparative analysis study of the nine polynomial regression models developed.

What needs to be documented ?

All the results of the above tasks must be present with your analysis.

Assignment 1-B Polynomial Regression and Regularization

Problem Statement

- The given dataset is a fish dataset consisting of ~200 data points, each having two feature variables and one continuous target variable.

Feature 1 : Width Of Fish **Feature 2 :** Height Of Fish **Target Feature :** Weight Of Fish

https://drive.google.com/file/d/1s6NZ0JXKlL2-3Qk5KlsX-FFVKJoGhPMu/view?usp=sharing

- **Task 1: Data Preprocessing**

  a. Load the shared dataset into a pandas DataFrame.

  b. Normalize the feature variables by utilizing the formula: X' = (X - μ) / σ where μ represents the mean of the feature column, and σ represents the standard deviation of the feature column.

  c. Some cells contain NAN / Null Values. Predict the values using the mean of the existing values of the corresponding feature.

  d. Shuffle the dataset and split the dataset into training and testing sets (80% for training and 20% for testing).

- **Task 2: Polynomial Regression**

a. Develop nine polynomial regression models (with degrees varying from 0 to 9) to predict the target variable based on the two input feature variables. Determine the degree of the polynomial which best fits the given data.

b. Build polynomial regression models of best fit degree(obtained from above) with the following generalized regularized error function:

$$\frac{1}{2}\sum_{n=1}^{N}\{t_n - \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q$$

Note : 1. Build different regression models by taking q as 0.5,1,2,4.

2. Experiment with λ values between 0 and 1 to obtain the optimal model for each value of q.

3. Use both Stochastic and Batch Gradient Descent. Report best models for each of the method.

- **Task 3: Graph Plotting**

a. Surface plots for the nine polynomial regression models and the four optimal regularized linear regression models.

- **Task 4: Comparative analysis**

a. Perform comparative analysis of the four optimal regularized linear regression models and the best-fit polynomial regression model among the nine models developed.

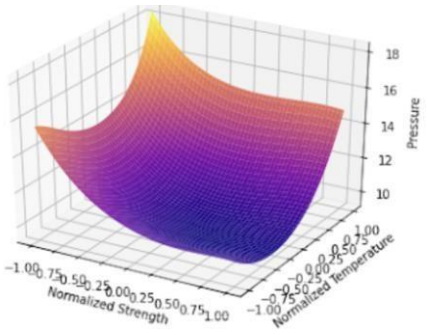
What needs to be documented ?

i) Give a brief description of your model, algorithms and how you implemented the regularization.

ii) Tabulate the training and testing errors obtained using polynomial regression models of various degrees and your observations on overfitting.

iii) Surface plots of the predicted polynomials (Plot of x_1, x_2 vs y(x_1, x_2) where y is the predicted polynomial.)

Example of a surface plot :



iv) The comparative analysis study of the four optimal regularized regression models and best-fit classic polynomial regression model.