IE343: Statistical Machine Learning

Assignment 1

Professor: Seyoung Yun

(TA : Mingyu Kim)

1. Instruction

Write a code for implementing the 'fit' function in linear regression class (`./Apps/linear_regressor.py`). Theoretically, this function is based on the least square method. The linear regressoion class has a member variable which represents "weights" in this model.

In this assignment, you need to consider which polynomial feature is appropriate for improving predictability of your model. Before making your regression model, you can take a look at some examples of polynomial feature methods given equally spaced X ($X = \{x : x \in [0, 1]\}$) (`./Apps/polynomial.py`). In order to use polynomial feature method, you need a predefined dimension. Accordingly, the report should suggest the dimensions of polynomial feature that best fit this dataset and explain why you chose it. The best dimension can be demonstrated by showing average mean squared error of the entire dataset and its standard deviation. Also, you visualize training and testing data and its fitted line together. (`./Apps/result_(i).png`)

The report includes an theoretical background for linear regression, a description of the model you've written and average root mean squared error and several plots within 1 page (Please write a report concisely). Also, you fill out the code in the attached python file completely.

2. Data

The provided data is generated by specific polynomial equation, but it is not released for students. The number of datasets is 10 and each dataset consists of 30 training data and 10 test data. In this assignment, the regression should be trained by only one of datasets. After trained your model, you compare real 'y' values to fitted values in terms of mean squared error(MSE) (By default, there is no need to consider how to manipulate dataset, you just run `./linear_regression.py`).

- The individual dataset has 30 training data and 10 test data.
- This data comes from a specific polynomial but it is not release.
- When you train and evaluate the model, you shoud use the same number of training and test data.
(There will be no problem if you use the given (data import) module).

3. Implementation

This code should be written using the 'Numpy' package. Not allowed to use other packages. (Scipy, Tensorflow, pytorch and etc.) You write down least square method ('fit' method) in (`./Apps/linear_regressor.py`) and (`./Apps/ridge regressor.py`) You can also manipulate the dimensions of polynomial features ('`feature_dimension`' variable in `./linear_regression.py`)

4. Question

  A. Write down the ordinary least squares estimator for "weights" and its setting

  B. Plot the train, test data and fitted line in any dataset. and explain why you choose the dimensionality of features with respect to data and your model.

  C. Write down the best average MSE and its deviation when you consider a total dataset.

  D. If you set a 9 dimensionality of features, what behavior of this model have in this dataset. Also, plot one graph among this dataset. Compare this model to the model you consider the best in terms of average MSE. and explain your opinion when you apply linear regression method to other dataset.