Foundations of Data Science (CS F320) Assignment 3

Submission by 11:59 PM on 15-Nov-2018

The Problem

Having a basic understanding of Machine Learning, Rabbit and Tortoise went to their professor for a project. The professor asked his TAs to suggest a task, for which they suggested implementing linear regression in different ways on a given dataset.

Dataset

The dataset and it's description is found here `https://archive.ics.uci.edu/ml/datasets/Combined+Cycle+Power+Plant`. Your task will be to predict the EP(power consumption) values and the loss over a test set. The test set in this case will be the last 2000 examples in this dataset, with the rest for training. You can use *pandas* to read data from the *.xlsx* file.

Part A: Normal Equations Method

You will be implementing the solution detailed in section 3.1.1 of Bishop. Refer to equations 3.15-3.17.

Part B: Gradient Descent Method

You will define the loss function to be half the sum of squared errors over the training set and minimize the loss function using Gradient Descent. Choose an appropriate learning rate and stopping criteria.

Part C: Gradient Descent with Regularization

As an extension to Part B, you will experiment with L1 and L2 regularization. Using a part of the training data for validation, determine the best value for the regularization coefficient.

Implementation Details

You will have to implement this in *python* using *numpy*, *pandas* and *matplotlib*. No other libraries are allowed.

Submission

Submission will be via CMS. Gather all your code into a single .py file. Along with that, make a document containing the following. Mention the loss over the test set for different methods. For method 3, make a plot showing the validation loss as a function of regularization coefficient. Comment on the effect of regularization on the regression coefficients. Zip the code and this document and name it with your ID numbers.

For Queries

Anirudh Srinivasan (2015A7PS0382H)
Phani Shankar Ede (2015B3A70420H)