Homework 02

ECE 469/568 – Machine Learning

SIU
Southern
Illinois
University
CARBONDALE

Date: 09/27/2024
Due data: 10/07/24
Section: Regression in Machine Learning
Instructions: Solutions must include Matlab codes, plots, and numerical results

[Q-1:] In this question, you are required to use a software package that supports machine learning, preferably Python-based Scikit-Learn libraries/classes.

(a) Download the dataset "datasetHW2P1.csv" from D2L. This dataset consists of inputs (x) to a non-linear system and the corresponding outputs (y). Split this dataset to create two sub datasets for training and testing.

(b) Use polynomial regression in machine learning to fit five models with following model complexities.

- polynomial with degree 1.
- polynomial with degree 2.
- polynomial with degree 3.
- polynomial with degree 4.
- polynomial with degree 5.

(c) Use 10-fold cross-validation to find the model that optimally fits to the given dataset. Plot the training, cross-validation, and testing errors against the model complexity (i.e., the degree of the polynomial).

(d) Consider a degree 4 polynomial as your model. Then use ridge-regression and find the best hyperparameter λ via 10-fold cross-validation. Plot the cross-validation error versus $\log_e(\lambda)$.

[Q-2:] In this question, you are required to use a software package that supports machine learning, preferably Python-based Scikit-Learn libraries/classes.

(a) Download the 'housing.csv' data-set from the following link and load it.
`https://github.com/ageron/data/tree/main/housing`

(b) Data preprocessing: You have already performed the following tasks for this dataset in Homework-01. Hence you may reuse your results in Homework-01 for the followings.

- Prepare the data by choosing the 'median_house_value' as the output and the rest as the input features.

- ‘ocean_proximity’ is a text attribute (categorical). You can either drop this feature or transform it into numerical values by using Scikit-Learn classes such as ‘OneHotEncodee’ or ‘OrdinalEncoder’
- Clean the data by either dropping the missing values or replacing the missing values with the median. (hint: use SimpleImputer class in SciKit-Learn)
- Carry out feature scaling either via normalization or standardization.
- Create a training data-set and a test data-set.

(c) Use linear regression to develop a machine learning model for prediction of ‘median_house_value’ for future inputs and analyze the test errors. Explicitly express the corresponding optimal weights and the final learned model. Use graphical illustrations where necessary to represent your results.

(d) Use cross-validation techniques improve the generalization of the model and analyze the root mean square error. Use graphical illustrations where necessary to represent your results.