IC 272 - Data Science III

Assignment 4: Regression

Deadline: October 20, 2024: 23.59 Hr.

Linear and Polynomial Regression

Dataset Description:

You are given a data file **abalone.csv** containing some structural information about marine snails, Abalones. Abalones are one of the key players in marine ecology and hence, studying them is important for maintaining a healthy marine environment. This dataset was prepared with the aim of making the age prediction of these snails easier. Customarily, the age of Abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope. But it is a tedious and time-consuming task. Therefore, other measurements, which are easier to obtain, are used to predict age.

This dataset contains those measurements of a snail based on which we can predict the number of its rings. As this task is predicting a continuous value given some measurements in continuous values; we pose this as a regression problem. and Following is a brief description of the collected data:

Independent variables/ Attributes/ Features:

i. "`Length`": Longest shell measurement in mm.

ii. " `Diameter`": The diameter of the shell calculated as perpendicular to length in mm.

iii. "`Height`": Height of the shell with meat in the shell in mm.

iv. "`Whole Weight`": Weight of the whole Abalone in gms.

v. "`Shucked Weight`": Weight of meat without the shell in gms.

vi. "`Viscera Weight`": Gut-weight (after bleeding) in gms.

vii. "`Shell Weight`": Weight of the shell after being dried in gms.

Dependent variable/ Target Attribute:

i. "`Rings`": Number of rings in a shell. (Adding 1.5 to the number of rings gives the

age of abalone in years).

Problem Statements:

I. Read the file **abalone.csv** using **Pandas** and split the data into train and test sets using the function from **scikit_learn** (use random_state=42). Train data should contain 70% of the tuples and test data contain the remaining 30% of the tuples. Save the train and test data as **abalone_train.csv** and **abalone_test.csv**.

II. Find the `attribute` that has the highest *Pearson correlation coefficient* with the target attribute `Rings`. Use this `attribute` as the input variable and build a simple linear (straight-line) regression model to predict the number of rings. Implement the **method** taught in the class. Do NOT use **any built-in** classification function.

i. Plot the best-fit line on the training data where the x-axis represents the chosen attribute value and the y-axis represents the number of rings, the target attribute.

ii. Find the prediction accuracy on the training data using root mean squared error.

iii. Find the prediction accuracy on the test data using root mean squared error.

iv. Draw a scatter plot of *actual* `Rings` (x-axis) vs *predicted* `Rings` (y-axis) on the test data.

III. Find the `attribute` that has the highest *Pearson correlation coefficient* with the target attribute `Rings`. Use this `attribute` as the input and build a simple nonlinear regression model using polynomial curve fitting to predict Rings. Implement the **method** taught in the class. Do NOT use **any built-in** classification function

i. Find the prediction accuracy on the training data for the different values of degree of the polynomial (**`p = 2, 3, 4, 5`**) using root mean squared error (RMSE). Plot the bar graph of RMSE (y-axis) vs different values of degree of the polynomial (x-axis).

ii. Find the prediction accuracy on the test data for the different values of degree of the polynomial (**`p = 2, 3, 4, 5`**) using root mean squared error (RMSE). Plot the bar graph of RMSE (y-axis) vs different values of degree of the polynomial (x-axis).

iii. Plot the best-fit curve using the best-fit model on the training data where the x-axis represents the chosen attribute value and the y-axis is Rings. (Note: The best-fit model is chosen based on the p-value for which the test RMSE is minimum.)

Auto-regression

Dataset Description:

You are given a data file **asianpaints.csv** containing the opening price of the Asian Paints stocks for the time period January, 2020 to April, 2021. Rows in this file are indexed with dates; the first column represents the date and the second column represents the opening price of the stock that day. We will use this dataset to build an autoregression (AR) model.

A general autoregression (AR) model estimates the unknown data values as a linear combination of given lagged data values. For example, data value at `(t+1)` instant, denoted by `x(t+1)`, can be estimated from its previous p instance values, such as `x(t+1) = w0 + w1*x(t) + w2*x(t-1) + … + wp*x(t-p+1)`. The coefficients `w0, w1, … wp` can be estimated by training the autoregression model on the training dataset.

I. Split the data into two parts. The initial 65% of the sequence is for training data and the remaining 35% of the sequence is for test data (*You may use slicing operation for the same to maintain the order of the sequence. Note that, you should not shuffle randomly*). Plot the train and test datasets.

II. Build an autoregression model with **one-day lag** using the train data and print the coefficients. Implement the **method** taught in the class. Do NOT use **any built-in** classification function.

III. Make a **one-step ahead** prediction for the test dataset and do the following (Implement the **method** taught in the class. Do NOT use **any built-in** classification function):

i. Draw a line plot showing actual and predicted test values.

ii. Compute and print RMSE (%) and MAPE (%) between the actual and predicted test data.