Week 3 - Regression

DS3010 - Introduction to Machine Learning

Instructions

1. Provide commented, indented code. Variables should have meaningful names.
2. Submit one .ipynb file containing all answers. The name should be [student name][roll_number] assignment[number].ipynb
3. Read the questions carefully before answering. If a question asks to follow a particular approach or to use a specific data structure, then it must be followed.
4. Write questions in separate text blocks in Jupyter Notebook before the code blocks containing answers.
5. All plots should have appropriate axis labels, titles, and legends.

Task for the Lab

1. Ordinary Least Square Method (9 points)

a. Randomly generate an independent variable 'x_i' of size 100 between 0 and 5. Afterwords generate dependent variable 'y_i', using equation $y_i = x_i**2 + ax_i + b + e_i$, where a, b and e_i are also randomly sampled. Where a and b are fixed for all x_i for and e_i varies. Generate 100 such x_i and y_i pairs. (1)

b. Find the linear regressor $y = mx+c$ which best fits the above generated data. Hint: Use the following formula to find parameter: (1)

$$\beta = (X^T X)^{-1} X^T y$$

c. Evaluate the model using mean square error (mse) (import from sk-learn) and print mse error. (1)

d. Print the learned parameter and plot data along with the learned fitted line. (1)
e. Report mse loss and learned parameters using linear regression model from sk-learn. Plot the data with the learned fitted line. (2)
f. Fit quadratic curve on same data using linear regression model from sk-learn. Report mse loss and learn Parameters. Plot new fitted curve on previous plot. Analyze the result. (3)

2. Multivariate Regression

a. **Data Preprocessing (7 points)**
   i. Load the given medical expenditure dataset for the regression task. (1)
   ii. Find all feature vectors with a null value and impute them using the appropriate method with justification. (2)
   iii. Find out the columns with the categorical values and convert them into numerical features using appropriate methods. (2)
   iv. Separate features and target column (charges). (1)
   v. Split the data into train-tests in a ratio 8:2. (1)

b. **Linear Regression (8 points)**
   i. Create an instance of a linear regression model. (1)
   ii. Fit the model to training data created in the previous question. (1)
   iii. Store the prediction for the training set. (1)
   iv. Find the prediction for the testing set. (1)
   v. Compute the mean squared error loss for the training set. (1)
   vi. Compute the R2 score of the model on the testing set. (1)
   vii. Plot bar graph of feature with its corresponding weight and print the top 2 important features. (2)

c. **Ridge Regression (8 points)**
   i. Create an instance of a ridge regression model with "alpha" = 1. (1)
   ii. Fit the model to training data created in question 2a. (1)
   iii. Store the prediction for the training set. (1)
   iv. Find the prediction for the testing set. (1)

v. Compute the mean squared error loss for the training set. (1)
vi. Compute the R2 score of the model on the testing set. (1)
vii. Plot bar graph of feature with its corresponding weight and print the top 2 important features. (2)