CSC 546/746

Assignment 3

(25 points)

1. There are 1000 rows of data in "hw03_data.txt" from 1000 companies. The first column is the amount of money that a company invested in research and product development (X). The second column is the profit of the company (y). You will use linear regression model to analyze the data.
2. (1 point) Download "hw03.ipynb" and "hw03_data.txt" from the Blackboard and upload (or copy) to your Jupyter Notebook folder. (Make sure you put the files in the same folder) Load hw03.ipynb project to Jupyter Notebook.
3. (2 points) The code for loading data from the file has been provided. Before initialize X and y, you need to scale the data with Scikit-Learn's StandardScaler class. (Check for the Preprocessing document for reference)
4. (2 points) Visualize the dataset. (Remember to adjust the range of X and y according to the dataset)
5. (7 points) Training the linear Regression model with Scikit-Learn's LinearRegression model:
    a. Splitting the dataset into the Training set (80%) and Test set (20%).
    b. Fit the training data to the model.
    c. Calculating the Intercept and the Coefficient. Print the result to the screen.
    d. Predict and print the Test set results.
    e. Calculate and print the R squared value.
6. (8 points) Solving the same problem with gradient descent algorithm:
    a. Training the model with training data.
    b. Calculating the Intercept and the Coefficient. Print the result to the screen.
    c. Predict and print the Test set results.
    d. Calculate and print the R squared value.
7. (5 points) The model should fit the data well, but the R squared value is close to 0.9, which is high but not as high as we expected.
    a. Analyze the reason.
    b. How to improve the R squared value?
    c. Create a new cell by the end of hw03.ipynb and provide the code to improve the R squared value.
8. Submit "hw03.ipynb" to the Blackboard.