HW: Linear Regression

ML4FS
Prof. Overdorf

Spring 2024

Overview

In this assignment, you will learn how to utilize scikit-learn's linear regression module to analyze the diabetes dataset. Linear regression is a fundamental technique used for predictive modeling when the relationship between the independent variables (features) and the dependent variable (target) is assumed to be linear.

Start by importing the relevant libraries:

```
import pandas as pd
from sklearn.datasets import load_diabetes
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
import matplotlib.pyplot as plt
```

Task 1: Dataset Exploration

In this task, you will start by loading the diabetes dataset from scikit-learn. This dataset comprises baseline variables such as age, sex, body mass index (BMI), average blood pressure, and six blood serum measurements for 442 diabetes patients. The target variable is a quantitative measure of disease progression one year after baseline.

After loading the dataset, explore its features, target variable, shape, and statistical summary to understand its characteristics.

```
# Here is some code to get you started
# It loads the data into a pandas data frame
diabetes = load_diabetes()
diabetes_df = pd.DataFrame(diabetes.data, columns=diabetes.feature_names)
diabetes_df['target'] = diabetes.target
...
```

Task 2: Data Preprocessing

Data preprocessing is essential to ensure the dataset is suitable for modeling. In this task, perform any necessary preprocessing steps, such as handling missing values and scaling the features. Decide what to do *if there are* any missing values in the dataset. Additionally, determine if the features need to be scaled to ensure that they are all on the same scale.

Task 3: Applying Linear Regression to a Single Feature

Apply linear regression to the preprocessed dataset to predict the target variable (disease progression) based on a single feature.

- Select Feature: Choose one feature from the dataset to use as the independent variable (feature) for the linear regression model. Select any feature.
- Split Data: Split the dataset into training and testing sets using the train_test_split function from scikit-learn. This ensures that the model's performance can be evaluated on unseen data.
- Train Model: Instantiate a linear regression model from scikit-learn's LinearRegression class and fit it to the training data.
- Predictions: Use the trained model to make predictions on the testing set.
- Evaluate Performance: Assess the performance of the model using appropriate evaluation metrics.

Task 4: Interpretation of Results

Interpret the coefficients obtained from the linear regression model. Visualize the regression line along with the original data points to understand the relationship between the selected feature and the target variable.

Task 5: Linear Regression with All Features

In this task, apply linear regression to the preprocessed dataset using all available features to predict the target variable (disease progression).

- Train the Model: Instantiate a linear regression model from scikit-learn's LinearRegression class and fit it to the entire dataset, including all available features.
- Predictions: Use the trained model to make predictions on the entire dataset.
- Evaluate Performance
- Interpret results