BIRZEIT UNIVERSITY

Electrical and Computer Engineering Department
Machine Learning and Data Science - ENCS5341
Assignment #2
Submission deadline: 23/11/2024

This assignment may be completed by a group of up to two students.

Machine Learning Project: Regression Analysis and Model Selection

Project Overview

The objective of this project is to build a series of regression models using a dataset, evaluate and compare their performance, and apply various techniques to improve model accuracy and prevent overfitting. The focus will be on both linear and nonlinear regression models. Students will also use feature selection methods and regularization techniques, followed by hyperparameter tuning, to select the optimal model. This project can be done **in groups of two students** at most.

Steps and Requirements

1. **Dataset**
   Cars Dataset: Dataset from online selling website yallamotors.
   https://www.kaggle.com/datasets/ahmedwaelnasef/cars-dataset/data

   This dataset, scraped from the YallaMotors website using Python and the Requests-HTML library, includes around 6,750 rows and 9 columns. It's well-suited for Exploratory Data Analysis (EDA) and machine learning tasks, particularly for predictive modeling using algorithms like Linear Regression. The main objective of this dataset is to predict car prices, making it ideal for developing regression models to understand the relationship between various features (e.g., car make, model, year, mileage, engine size, etc.) and the target variable (car price).

   Through EDA, you can explore patterns, outliers, and relationships in the data, which will help refine your model. For the machine learning task, Linear Regression could be a good starting point, but more complex regression models can also be applied if necessary to capture non-linear relationships and improve predictive accuracy.

   Note that car prices are listed in various currencies. To ensure consistency, you may need to standardize all prices to a common currency, such as USD, for a uniform target variable. This will help avoid discrepancies and improve the accuracy of any predictive modeling.

   Data Preprocessing Steps:

   - Clean the dataset by handling missing values, encoding categorical features, and normalizing or standardizing numerical features if necessary.
   - Split the dataset into training, validation, and test sets. A common split would be 60% for training, 20% for validation, and 20% for testing.

2. **Building Regression Models** Implement a set of linear and nonlinear regression models:
   - **Linear Models**: Linear Regression, LASSO (L1 Regularization), Ridge Regression (L2 Regularization). *See the details in point 5 below.*
   - **Use the closed-form solution:** Apply the closed-form solution for the linear regression model to solve the system of linear equations and obtain the model parameters. Then, compare this model with the one derived using the gradient descent method. Implement this part without using any external APIs or libraries for linear regression.
   - **Nonlinear Models**: Polynomial Regression (vary the polynomial degree from 2 to 10), and standard Gaussian kernel, known here as a Radial Basis Function (RBF).
3. **Model Selection Using Validation Set** Evaluate each model on the validation set and use metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and R-squared to compare their performance. The best model can be selected based on the lowest MSE or highest R-squared score on the validation set.
4. **Feature Selection with Forward Selection** Use a **forward selection method** to iteratively add features to the model, selecting features that improve model performance. The forward selection process will:
   - Start with an empty model and gradually add features one at a time.
   - At each step, add the feature that, when included, minimizes the error on the validation set.
   - Stop once additional features no longer improve the model performance or a maximum number of features is reached.
5. **Applying Regularization Techniques** Use **LASSO and Ridge regularization** to control overfitting. These techniques will help to reduce the model complexity by penalizing large coefficients and potentially zeroing out less relevant features (LASSO).

   Steps:

   - Implement LASSO and Ridge regression with different values of the regularization parameter λ.
   - Use **Grid Search** to find the optimal λ value that minimizes the error on the validation set.
6. **Hyperparameter Tuning with Grid Search** Apply **grid search** to find the best hyperparameters for each model (e.g., λ for regularized models). This step ensures that each model is tuned for optimal performance.
7. **Model Evaluation on Test Set** After selecting the best model based on the validation set, evaluate the chosen model on the test set to obtain a final performance metric. Report on how well the model generalizes to unseen data.
8. **Optional:** Try identifying another relevant target variable in the dataset and build a regression model to predict its values.
9. **Reporting the Results** Prepare a detailed report on the findings, including:
   - Description of the dataset, preprocessing steps, and features used.
   - Details of each regression model and its performance on the validation set.
   - Explanation of feature selection results using forward selection.
   - Regularization results with the optimal λ values for LASSO and Ridge.
   - Model selection process with grid search and hyperparameter tuning.
   - Final evaluation on the test set and a discussion of the selected model's performance and limitations.
   - Visualizations to support findings (e.g., feature importances, error distribution, and model predictions vs. actual values).

Submission Requirements

- Compress the project files (code and report) into a single ZIP file.
- Submit both code and report files separately, following the provided file naming format.
- Adhere to the submission deadline and note the late submission policy.

Group Work Policy

This project can be completed in groups of **two students at most**. Each student should clearly state their contribution to the project in the report.

Submission:

A- A comprehensive report that describes the dataset and summarizes and discusses all the results and findings as required above.

B- Your code (python code) in either .py format or a Jupyter Notebook with both the code and visualizations.

C- Please compress your files, including both the code and the report, into a single zip file and submit it to the ritaj before the deadline. The file name should follow this format: "LastName_ID_Student1_LastName_ID_Student2.ZIP".

D- Late submissions will be accepted up to 3 days after the deadline, with a 10% deduction for each day delayed.

Hint:

```
You can use the following python libraries: Pandas,
NumPy, Matplotlib, Seaborn, sklearn
```