2023-CSE5ML-T4-W Machine Learning

Assessment 1: Design regression models

La Trobe University

Student: Benjamin Kereopa-Yorke

Student No: 21340711

Date: 13.08.2023

Table of Contents

Introduction

Leveraging data from the World Health Organisation (WHO), this study aimed to predict life expectancy based on a variety of health-related indicators.

My approach was systematic: I began by preprocessing the dataset, addressing missing values, and encoding categorical variables. Subsequently, the data was split into training and testing sets, ensuring a robust foundation for model evaluation. Two regression models, Linear Regression and Support Vector Machines (SVM) Regression were employed to decipher the intricate relationships between the health metrics and life expectancy.

Linear regression is a fundamental statistical and machine learning method used to predict a quantitative response variable based on one or more predictor variables. The main idea is to establish a linear relationship between the predictors and the target. In mathematical terms, it fits a linear equation that minimizes the sum of the squared differences between the observed values and the values predicted by the model.

Support Vector Machines (SVM) is a versatile algorithm primarily known for classification tasks but can also be applied for regression, known as Support Vector Regression (SVR). In SVR, the goal is to find a hyperplane that best fits the data, but instead of trying to maximize the margin between two classes (as in classification), SVR tries to capture data points within a specified margin error. The flexibility of SVMs lies in their ability to use different kernel functions, allowing them to model non-linear relationships.

Through rigorous training, evaluation, and hyperparameter tuning utilising a Google Colaboratory hosted Jupyter Notebook, I aimed to harness the best predictive performance from my models. This report explains my journey through this analytical process and the insights derived therein.

Task 1: Preparing the Data

Dataset Overview (Prior to Preprocessing):

The initial dataset, sourced from the World Health Organisation (WHO), comprised 2,928 entries spanning 20 columns. Each entry encapsulated a country's health indicators and the corresponding life expectancy. The dataset's variables were predominantly numerical, with 16 columns of type float64, 3 columns of type int64, and one categorical column (Status). Some columns, specifically Hepatitis B, Polio, and Diphtheria, had missing values, with each missing 19 entries.

Data Preprocessing Steps:

Handling Missing Values: The columns with missing values (Hepatitis B, Polio, and Diphtheria) were addressed by replacing the absent entries with the median of the respective columns. The choice of the median, as opposed to the mean, was influenced by its robustness to outliers, ensuring the imputed values did not skew the dataset.

Handling Categorical Variables: The dataset contained a categorical column, Status, indicating whether a country is "Developed" or "Developing". To make this data amenable for machine learning algorithms, the column was transformed using one-hot encoding. This process resulted in a binary column named 'Status_Developing', where a value of 1 indicates a developing country and 0 otherwise.

Data Splitting: To facilitate model training and evaluation, the dataset was divided into two sets: training and testing. 90% of the data was allocated for training, while the remaining 10% was reserved for testing. This split ensures a comprehensive training phase while retaining a portion of the data to evaluate the model's performance on unseen data.

Normalisation: To ensure that no feature disproportionately influenced the model due to its scale, normalisation was applied to the predictor variables X of both the training and test sets. This process transformed the data such that each feature's values lie between 0 and 1. Normalisation is crucial, especially for algorithms like SVM, which are sensitive to feature scales.

Dataset Description (After Preprocessing):

Post preprocessing, the dataset evolved to contain 2,928 entries with an additional column (Status_Developing), making the total 21 columns. The dataset was devoid of missing values, ensuring a complete data set ready for model training and evaluation.

Data preprocessing is a pivotal step in any machine learning pipeline. Raw data, while rich in information, often comes with inconsistencies, missing values, and variables of varying scales and types. Addressing these issues ensures that the machine learning algorithms can decipher the underlying patterns more effectively. The steps taken, from imputation to normalisation, were all geared towards refining the dataset, making it a fertile ground for accurate and insightful model training.

Task 2: Building & Assessing Regression Models

After preprocessing the dataset and establishing the training and test sets, I trained and evaluated two regression models: Linear Regression and SVM Regression. Both models were assessed using their default parameter settings.

Results with Default Parameters:

Model	Cross-Validation Score (Default)	Test Score (Default)
Linear Regression	0.830	0.862
SVM Regression	0.850	N/A

The SVM regression model slightly outperformed the linear regression model based on the average cross-validation scores. While both models provided decent performance, the inherent ability of SVM to capture non-linear relationships might be advantageous for this dataset.

To further enhance the performance of the models, I undertook a parameter finetuning process. For SVM Regression, I employed grid search, an exhaustive search method that tests a specified set of hyperparameters to identify the combination that provides the best performance.

For SVM Regression, I explored a combination of hyperparameters:

- C: Regularization parameter, values explored were [0.1, 1, 10, 100].
- kernel: Specifies the kernel type, values explored were ['linear', 'rbf', 'poly'].
- gamma: Kernel coefficient, values explored were ['scale', 'auto'].

Results Before and After Finetuning:

Model	Cross-Validation Score (Default)	Cross-Validation Score (Optimised)
Linear Regression	0.830	N/A (no tuning)
SVM Regression	0.850	0.926

Post finetuning, the SVM Regression model's performance was notably enhanced, showcasing the benefits of hyperparameter optimisation.

Lastly, I evaluated my optimised models on the test set to gauge their real-world predictive performance.

Model	Test Score (Optimised)
Linear Regression	0.862
SVM Regression	0.954

The SVM Regression model, especially post-optimisation, exhibited superior performance compared to the Linear Regression model on the test set.

Conclusion

My analysis revealed the nuanced capabilities of both Linear Regression and SVM Regression models. While both models showcased strong performance, the SVM Regression model, post-optimisation, won out. This underscores the importance of model selection and hyperparameter optimisation in machine learning tasks.