CT5165 Principles of Machine Learning: Assignment 3

Jiarong Li 20230033, 1MAI1

The School of Computer Science
University of Galway
j.li11@nuigalway.ie

1 Introduction

The goal of this assignment is to gain practical experience of performing regression on a real world dataset, using a machine learning package of your choice. To complete this assignment, you will train two different regression models and prepare a short report. This report will describe the methodology followed, and analyse the performance of the models that were trained.

2 Methodology

2.1 ML package for the regression task.

I apply scikit-learn library in this assignment as scikit-learn is one of the most robust machine learning libraries in python [1] and it provides different options for numeric calculations and statistical modelling which is powerful for training processing and making predictions, besides, the Linear Regression is one of its modet important sub-models [4]. Hence, I apply scikit-learn package for this regression task.

2.2 Dataset

I use cross-validation in this regression task to minimise the probability of overfitting. Overfitting happens when a model learns the detail and noise in the training data to the extent that it negatively impacts the performance of the model on new data [2]. And k-fold cross-validation is one of the most powerful techniques to limit overfitting. It is easy to implement, easy to understand, and results in skill estimates generally have a lower bias than other methods [3]. I apply k-fold cross-validation by randomly dividing the training dataset into 5 folds (4 folds for the training dataset and one fold for the validation dataset, and randomly repeating 5 times) for the training process. During the cross-validation process, I also apply REF and GridSearchCV library to select the optimal number of features and apply it in the final model, then obtain the final hypothesis trained with the whole training dataset.

2.3 Models: Select two different regression algorithms that you will apply to the dataset to learn two different regression models.

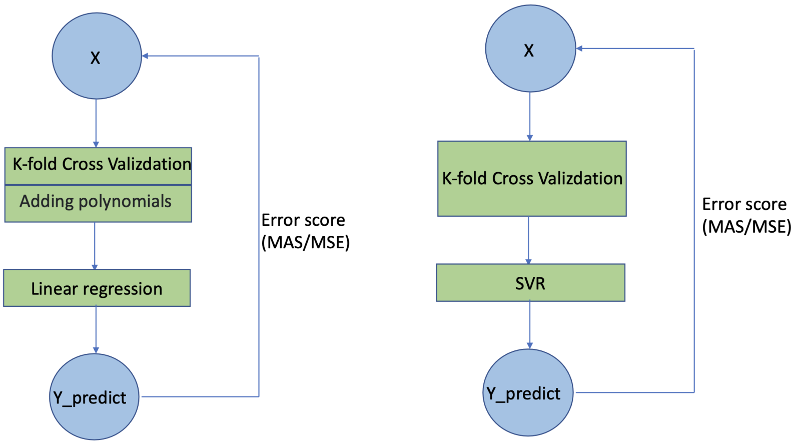
I use linear regression submodule and support vector regression submodule of sckit-learn package to learn two different models.

Linear regression is a supervised machine learning algorithm generally applied in regression tasks. Linear regression makes the prediction of the dependent variable based on the independent variables by finding the linear relationship between the dependent variables and independent variables. It is simple to implement and has less complexity compared with other algorithms. On the other hand, linear regression may lead to over-fitting but it can be avoided using some techniques such as cross-validation [5].

Linear regression finds the linear relationship between dependent and independent variables. However, it can over-simplifies real-world problems. Support Vector Regression takes into account not only the essential linear separation between two classes, but also consider the situation in multidimensional space by finding a line or a hyperplane to separate the two classes which can give us a high prediction accuracy. It gives us the flexibility to define how much error is acceptable in our model and will find an appropriate line (or hyperplane in higher dimensions) to fit the data [6]. But they do not perform very well when the data set has more noisy and not suitable for large datasets [5].

2.4 Describe the process you followed while developing each model.

The design of the linear regression model and SVR model is shown in Fig 1. The process for developing each model is: (1) environment setting, (2) data preprocessing, data loading (3) cross-validation training, (4) model design, (5) model training and testing, (6) and making predictions.



(a) Linear regression model architecture.

(b) SVR model architectuer.

Figure 1: Linear regression model architecture and SVR model architecture.

I import the necessary libraries such as numpy, pandas, matplotlib, and I also import the libraries from the sklearn package to help me build linear a regression model and SVR model for this regression task.

In data processing, I write an assistant data processing function to prepare the training and testing data. I first convert the CSV file to readable data frame format. Then, I do data pre-processing for the independent variables such as dividing ”datetime” variable into ”year”, ”month” and ”time” three variables to make the independent variables informative. I apply OrdinalEncoder method for independent variables such as ”time” and ”var2” to make categorical attributes continuous. Then, I divided the train data set into train data with independent variables and train data only with target labels.

I apply the k-fold cross-validation method to estimate whether the model is expected to perform with a written assistant function with a built-in method from sklearn package. I also apply GridSearchCV from sklearn package to find the optimal number of features in the data set and the results Fig 2 show that we can use all the independent variables from the data set. Hence, the hyperparameter setting for the number of features in two models is the number of the independent variables in the dataset.

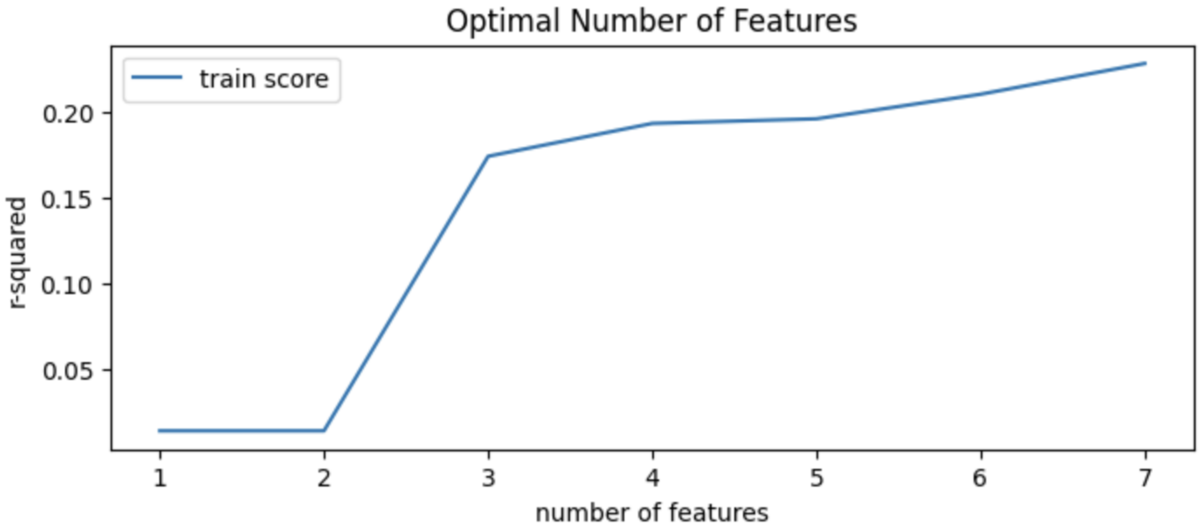


Figure 2: Optimal number of features.

As aforementioned 2.3, I apply the linear regression algorithm and support vector regression algorithm in model designing. I apply 5 epochs train for each model (linear regression model and SVR model) with the whole training data to obtain the trained linear regression model and obtain the trained SVR model. Then, I make predictions of the test data set with the trained models and save the result to a new CSV file.

3 Evaluation

I apply two approaches to evaluate these two designed models: (1) calculate error as measures in the training process with RMSE metrics, MAE metrics and R2 metrics, and (2) plot the predict and ground truth target.

The average error for the linear regression model with RMSE, MAE, and R2 metrics evaluations are shown in Table 1. The Fig 3 visualises the results by showing us the relation between the target and predicted label, the plot of train data with train targets and the plot of test data with predicted labels.

Table 1: Linear regression model evaluation with metrics

Metrics	score and error
average r2 score	0.4821
average mean absolute error	54.7154
average mean squared error	6042.9550

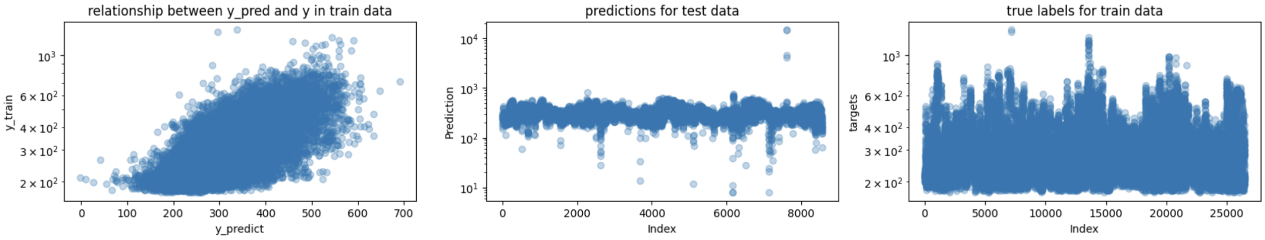


Figure 3: Visualise the results of linear regression model.

The average error for the SVR model with RMSE, MAE, and R2 metrics evaluations are shown in Table 2. The R2 score is negative in this case, which might mean that the chosen model SVR does not follow the trend of the training data set, or there is noise in the data set or the size of the data set is not properer. The Fig 4 visualises the results by showing us the relation between the target and predicted label, the plot of train data with train targets and the plot of test data with predicted labels.

Table 2: SVR model evaluation with metrics

Metrics	score and error
average r2 score	-0.0570
average mean absolute error	74.5757
average mean squared error	12333.1589

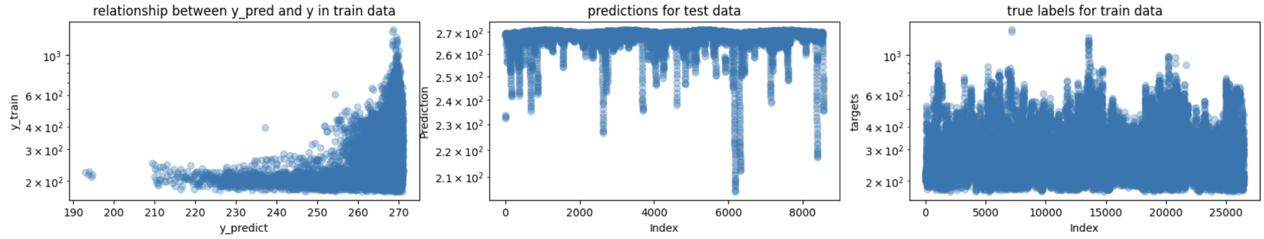


Figure 4: Visualise the results of linear regression model.

References

[1] Avijeet Biswal. Sklearn linear regression, 2022. URL https://www.simplilearn.com/tutorials/scikit-learn-tutorial/sklearn-linear-regression-with-examples.

[2] Jason Brownlee. Overfitting and underfitting with machine learning algorithms, 2019. URL https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/.

[3] Jason Brownlee. A gentle introduction to k-fold cross-validation, 2020. URL https://machinelearningmastery.com/k-fold-cross-validation/.

[4] Damian Ejlli. Five regression python modules that every data scientist must know, 2021. URL https://towardsdatascience.com/five-regression-python-modules-that-every-data-scientist-must-know-a4e03a886853.

[5] Gaurav Sharma. 5 regression algorithms you should know – introductory guide!, 2021. URL https://www.analyticsvidhya.com/blog/2021/05/5-regression-algorithms-you-should-know-introductory-guide/.

[6] Tom Sharp. An introduction to support vector regression (svr), 2020. URL https://towardsdatascience.com/an-introduction-to-support-vector-regression-svr-a3ebc1672c2.