COMP1816 - Machine Learning Coursework Report

Bahadir Erkam Bakoglu - 001089837

Word Count: 1689 words

1. Introduction

This paper describes the pre-processing and application phases of multiple models in regression and classification cases and compares their efficiency results respectively. Univariate linear regression and kernel ridge regression are used as algorithms for the regression task. The algorithm implemented as the baseline model was univariate linear regression algorithm and kernel ridge regression was the additional model for comparison for efficiency. After comparing the baseline model with the comparison model, the results indicated that the univariate linear model was slightly more efficient. As for the classification task, the decision tree algorithm and the support vector classifier (SVM) are used as algorithms. The algorithm implemented as the baseline model was the decision tree algorithm and the additional algorithm for efficiency comparison was the support vector classifier. Comparing the baseline model with the comparison model revealed that SVM performed slightly better than decision tree algorithm.

2. Regression

2.1. Pre-processing

The data set for the regression task included values related to housing in California. The columns of the data set are No., longitude, latitude, housing median age, total rooms, total bedrooms, population, households, median income, median house value and ocean proximity. Since the objective of the task is to predict median house value, it is essential to remove the unrelated data and outliers for better accuracy.

The preparation for the pre-processing begins with combining the two separate testing and training data sets into a single data set to allow for an easier pre-processing phase. After the preparation, it is possible to upload the data set and determine the necessary features for the prediction of the target. It is clear that No., longitude and latitude are not related to median house value based on the fact that No. is a unique value, and longitude and latitude are coordinates that are not linked to the target, median house value, so it is reasonable to drop these features. The next step is to inspect the details of the data in the remaining features. By looking at the overall count of the entries for every column it is clear that the data set includes lines that have empty values. A possible solution in this scenario is to remove the lines that have empty values since there are only several of such lines and the impact of removal has minimal effect on the final prediction. Then, it is important to separate numerical and categorical values in order to deal with the outliers. After displaying the numerical details, it is apparent that there are numerous outliers that have to be managed. Every feature can be displayed as a histogram along with the minimum and maximum values. The outliers can be removed by limiting the minimum and maximum values of a feature. The process is applied for every feature until the outliers are removed.

2.2. Methodology

The main model implemented in the regression task is kernel ridge regression. The kernel ridge regression is used in this scenario since it is capable of computing the mean squared error and classification error rate based on the results of a single training. This is possible because the hyper parameters can be optimally adjusted. In addition, it is possible to compute with a single training the solutions corresponding to multiple values of the ridge if an implementation with singular value decomposition is made. This makes kernel ridge regression a fast and versatile model to be implemented in this scenario.

2.3. Experiments

2.3.1. Experimental Settings

The baseline model implemented for the regression task is univariate linear regression. Despite its simplicity, multivariate linear regression is a very useful algorithm that can be implemented quickly and easily to provide satisfactory results. It is also possible to train univariate linear regression efficiently and easily. The hyper parameter tuning of the models includes training and testing sizes of the data according to the pre-processed data set. The initial division of the training and testing sizes is 80 to 20 percent. However, this value is optimized to 85 to 15 percent to maximize accuracy. In addition, a learning rate of theta 0: 41075.503248247755 and theta 1: [42343.37535385] has been acquired.

2.3.2. Results

The achieved results of the baseline model and the main comparison model, univariate linear regression and kernel ridge regression, are evaluated by two values MSE (Mean Squared Error) for efficiency and R2 (Relative to square error) score for accuracy of the model. The reason MSE is used as an evaluation metric in this scenario is because of its differentiable and easier to optimize structure. As for R2, it measures the strength of the relation between the model and the data set. In addition, both MSE and R2 are easy to implement and optimize in linear regression tasks. The mean squared error of univariate linear regression model recorded as 7853511232.6919 and the R2 score is 0.5014. As for kernel ridge regression model, the mean squared error is 8329636025.7743 and the R2 score is 0.4712.

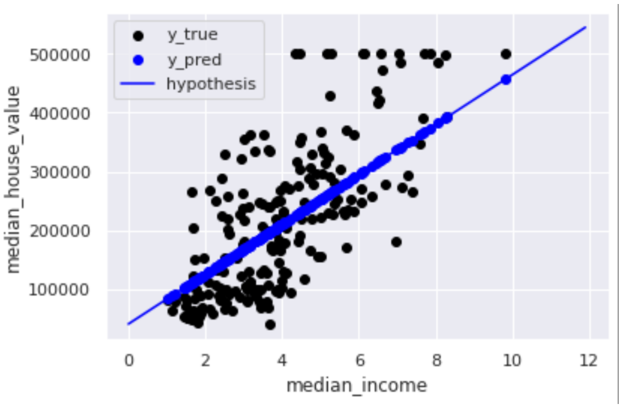

Figure 1. Univariate linear regression graph

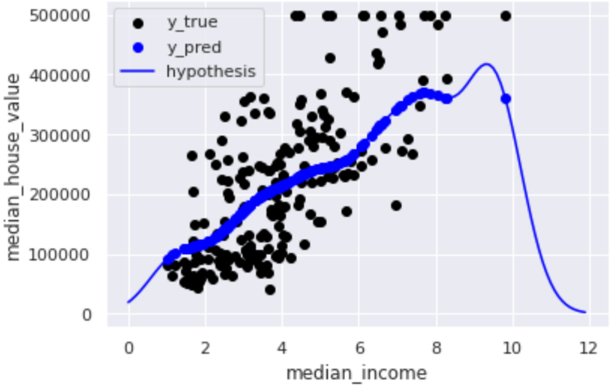

Figure 2. Kernel ridge regression graph

2.3.3. Discussion

As a result of the evaluation metrics, it is evident that the baseline model outperforms the main comparison model slightly. The reason for the difference between the models is mostly based on the nature of the algorithms. This is because the univariate linear regression performs better with less number of features while kernel ridge regression is a better choice in larger data sets with more features.

3. Classification

3.1. Pre-processing

The data set for the regression task included values related to the Titanic passenger details. The columns of the data set are PassengerId, Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Embarked and Target: Survived. Since the objective of the task is to predict what sorts of people were more likely to survive, it is essential to remove the unrelated data and outliers for better accuracy.

The preparation for the pre-processing begins with combining the two separate testing and training data sets into a single data set to allow for an easier pre-processing phase. After the preparation, it is possible to upload the data set and determine the necessary features for the prediction of the target. It can be observed that PassengerId, Name, Ticket, Fare are unrelated data since they are unique to every person and provide no value for the prediction model. In addition, SibSp and Parch are related to family relations and are not necessary for the prediction of the task. After removing the unrelated data, we can determine if there are any mistypes or empty data since there are features that are categorical. This can be done by listing unique values in Pclass, Sex, Embarked, Target: Survived. As observed there appears to be any mistypes however, there are lines with empty data. It is possible to remove the lines with the empty data considering the minority of the numbers of the lines with empty data. After that, by listing the numerical and categorical features in separate data sets it is possible to determine and remove the outliers in the data set. The features of the data set can be displayed in a histogram for visualizing the values in them. Then the values of the features can be limited so the outliers are removed. After cleaning up the data set, it needs to be converted into NumPy arrays and standardized for the use of classification models.

3.2. Methodology

The main model implemented for the classification task is SVM. SVM is used in this scenario because of the nature of the algorithm. A class separation margin between classes is necessary for SVM to work effectively. The data set includes many categorical features. This makes SVM an acceptable option for this task.

3.3. Experiments

3.3.1. Experimental Settings

The baseline model implemented for the regression task is decision tree algorithm. It has been implemented because the decision tree algorithm is a model that enhances decision making and predicting processes. In addition, because the inputs used in this machine learning algorithm do not directly affect the outcome. Rather, the outcome is predicted by examining the relationship between the different inputs. The initial division of the training and testing sizes is 80 to 20 percent. However, this value is optimized to 85 to 15 percent to maximize accuracy. In addition, minimum samples split is 2, minimum samples leaf is 9 are set as the hyper parameters.

3.3.2. Results

The achieved results of the baseline model and the main comparison model, decision tree algorithm and SVM, are evaluated by confusion matrices. In the classification tasks, the accuracy alone for is not reliable enough for predictions. As a result, the confusion matrix is an optimal evaluation method for classification tasks which result in two or more classes. This is because it is easier to gain a better understanding of what the classification model performs well and what types of errors it makes while maintaining an equal number of observations and establishing a more reliable output. The accuracy of the decision tree model was recorded as 0.6788 and the confusion matrix is [[0.29015544 0.12953368][0.19170984 0.38860104]]. On the other hand, for the SVM model, the accuracy is 0.6995 and the confusion matrix is [[0.21243523 0.20725389][0.09326425 0.48704663]].

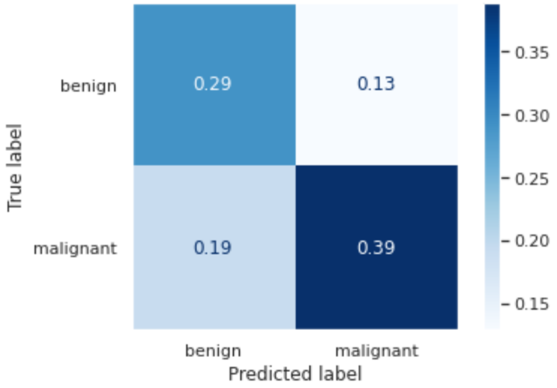

Figure 3. Decision tree confusion matrix

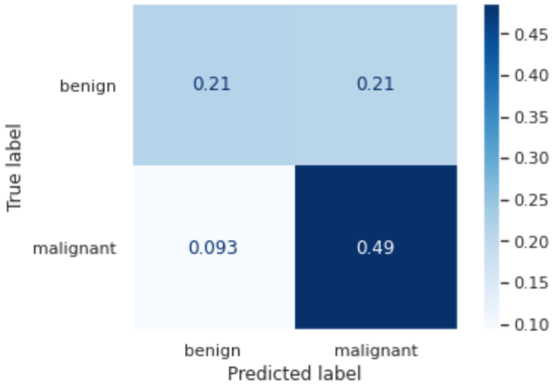

Figure 4. Support vector classifier confusion matrix

3.3.3. Discussion

According to the evaluation metrics, it can be observed that the main comparison model performs better than the baseline model. It is primarily the algorithmic nature of the models that are responsible for the differences between the results. SVM uses kernel to solve non-linear problems, whereas decision trees use hyper-rectangles in input space to solve them. This causes the decision tree model to be not as efficient as the SVM model in the case of non-linear tasks.

4. Conclusion

Therefore, it can be concluded that the baseline model, univariate linear regression, performed better than the main comparison model, kernel ridge regression. Meanwhile, the main comparison model, the SVM model, outperformed the baseline model, the decision tree algorithm, in the classification task. However, it is certain that the results can be perfected further. Such as, the accuracy and reliability of the results of both tasks can be improved by cleaning the data set further. In addition, considering most of the values in the data sets is in numerical format, it is possible to impute missing data instead of deleting the missing rows. This may lead to a slight increase in the accuracy and reliability of results.