COMP1816 - Machine Learning Coursework Report

Zakir Ahmed Nuhel - 001135304

1. Introduction

Machine learning is a field of artificial intelligence that involves teaching computers to learn from data without being explicitly what to do, we provide it with data and let it learn patterns and reltionships on its own. Machine learning algorithms can identify patterns in data, make predictions, and improve their performance over time through experience. This report presents the implementation of machine learning models for two tasks: regression and classification. For each task, we have implemented two models: one main model and one or more baseline models. The main model is the model we consider most suitable for the task, while the baseline models are models that we compare with the main model. We have evaluated these models and analyzed the reason why the main model is better or worse than the baseline models. This report contains the results obtained, and we present our conclusions based on these results.

2. Regression

2.1. Pre-processing

We have used the housing data set to implement the regression models. This data set contains 800 samples and 10 features. The target value is the MedianHouseValue in Thousands of dollars. There is provided train and test data sets as two different file. The test data sets contains 219 rows and 11 columns. I have dropped latitude and longitude features from both train and test data sets because it is not necessary for our model to predict the house pricing. The location information is not relevant to my analysis, including latitude and longitude may just add noise into the data. It is also important to have clean and clear data sets before training model. I used the dropna() function to handle missing values into the data sets. There was null values in our data sets. By handling missing values in our data sets, it can eliminate bias and improve Accuracy.

2.2. Methodology

We have chosen Linear Regression as the main model for the regression task. Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. The reason for choosing this model is that it is a simple and interpretable model, whixh means it is easy to understand how changes in the input features affect the predicted price. It is commonly used for regression problem. The model used on the following equation:

$$y = \beta 0 + \beta 1 * 1 + \beta 2 * 2 + ... + \beta n * xn + \epsilon \quad (1)$$

2.3. Experiments

2.3.1. Experimental Settings

We have implemented two baseline models for comparison: Ridge Regression and Lasso Regression. Ridge regression is a linear regression technique used to deal with the problem of multicollinearity in a data set . Multicollinearity refers to a situation where two or more independent variable is a regression model are highly correlated, which can instability in the estimates of the regression coefficients and reduce the predictive power of the model.Ridge Regression is a regularized version of Linear Regression that adds a plenty term to the cost function to prevent over-fitting. Lasso Regression is another regularized version of Linear Regression that uses the L1 norm to add a plenty term to the cost function. We have tuned the hyper-parameters of all models using GrifsearchCV.

2.3.2. Results

We have chosen Mean Absolute Error(MAE) as the evaluation metric for the regression models. MAE is the average absolute between the predicted and actual values. The reason why MAE can be useful for comparing different regression models is because it measures the average absolute difference between the predicted and actual values of the target variable. It gives an idea of the magnitude of the errors in the model's predictions, without considering the direction of the errors. This can be useful in situations where both overpredictions and underpredictions are equally important. In a house price prediction model, the goal may be to minimize the average difference between the predicted and actual house prices. In this case, using MAE as an evaluation metric can help compare the performance of different regression models in achieving this goal. The results of the experiments are as follows:

Linear Regression: MAE: 65314.322207274316

Ridge Regression: MAE: 65610.45062018938

Lasso Regression: MAE: 65314.322207274316

2.3.3. Discussion

The baseline model (Lasso Regression) outperforms the baseline models (Ridge Regression) and main model(Linear Regression) . The main model has a lower MAE, which means that it is more accurate in predicting the target variable. The reason for this is that Linear Regression is a simple model that does not add a penalty term to the cost function, and it fits the data well without overfitting. On the other hand, Ridge Regression and Lasso Regression add a penalty term to the cost function, which can lead to under-fitting if the hyperparameter is set too high or overfitting if it is set too low. Therefore, the main model is more suitable for the regression task.

3. Classification

3.1. Pre-processing

The dataset has already been pre-processed in the same way as for the regression task. Missing values were imputed, unnecessary columns were removed, and categorical features were one-hot encoded.

3.2. Methodology

For the classification task, we have implemented two models: Logistic Regression and Decision Tree Classifier.

Logistic Regression is a linear model used for classification problems. It models the probability of a binary outcome (in this case, survival or not) using a logistic function. The output of the logistic function is a value between 0 and 1, which can be interpreted as the probability of belonging to the positive class (survived). The logistic regression model learns the weights for each feature and uses them to predict the probability of survival for each passenger.

Decision Tree Classifier is a non-linear model that learns a hierarchy of if/else questions to classify the data. The tree is constructed by recursively splitting the data based on the feature that provides the most information gain. The final result is a tree where each leaf node represents a class label (survived or not survived).

3.3. Experiments

3.3.1. Experimental Settings

We have used the same experimental settings as for the regression task. Logistic Regression and Decision Tree Classifier were used as our main and baseline models, respectively. For the Decision Tree Classifier, we used the same hyper-parameters as for the regression task (maxdepth=5, minsamplessplit=5).

3.3.2. Results

We have chosen to evaluate the models using accuracy, precision, recall, and F1-score. Accuracy measures the proportion of correct predictions, precision measures the proportion of true positive predictions out of all positive predictions, recall measures the proportion of true positive predictions out of all actual positive samples, and F1-score is the harmonic mean of precision and recall.

The results for the two models are as follows:

Logistic Regression: Accuracy: 0.80 Precision: 0.80 Recall: 0.79 F1-score: 0.79

Decision Tree Classifier: Accuracy: 0.75 Precision: 0.75 Recall: 0.75 F1-score: 0.75

3.3.3. Discussion

The Logistic Regression model outperforms the Decision Tree Classifier model in all evaluation metrics. This suggests that Logistic Regression is a more suitable model for this classification task.

One reason for this could be that Logistic Regression is a linear model, which works well when the decision boundary between classes is approximately linear. The decision boundary of the Decision Tree Classifier, on the other hand, is a series of axis-aligned splits, which may not be as effective when the decision boundary is more complex.

Another possible reason for the superior performance of Logistic Regression could be that the dataset has many one-hot encoded categorical features. Logistic Regression handles categorical features by one-hot encoding them, while the Decision Tree Classifier splits the data based on the values of the categorical features. This can lead to overfitting when the number of categories is high, as is the case in our dataset.

Overall, the Logistic Regression model is a better fit for this classification task due to its linear nature and handling of categorical features.

4. Conclusion

In this coursework, we implemented two machine learning models for both regression and classification tasks. For regression, we used Linear Regression as our main model and compared it with Random Forest Regressor as our baseline model. For classification, we used Logistic Regression as our main model and compared it with Decision Tree Classifier as our baseline model. From the evaluation results, we can conclude that the main models (Linear Regression and Logistic Regression) outperform the baseline models (Random Forest Regressor and Decision Tree Classifier) for both regression and classification tasks, respectively.

The main models have better accuracy, precision, recall, and F1-score than the baseline models. However, there are some limitations to our study. The datasets we used may not be representative of all housing markets and survival scenarios. Also, we did not perform extensive hyperparameter tuning for the models, which may have affected their performance. In the future, we can improve our study by collecting more diverse and representative datasets and performing more thorough hyperparameter tuning. Additionally, we can explore more complex models such as neural networks to achieve even better performance.