Danmarks Tekniske Universitet

DTU

Case 1

Authors

02582 Computational Data Analysis
Maria Kokali - s232486
Marios-Dimitrios Lianos - s233558

March 19, 2024

Contents

1 Introduction/Data Description

This report presents our approach to build a predictive model that can estimate new values of a response variable with as much accuracy as possible.

We are going to perform techniques for model selection, validation and assessment on a dataset which consists of 100 observations (y,x), of response Y (vector), features X (100-dimensional feature matrix). Subsequently, we have 1000 additional observations on which we are going to do predictions based on the final model that we will choose.

In short, we are going to build a predictive model of Y based on X, by using Python.

Data Description

By observing our dataset, we see that y variable (float) is our target variable and the variables $X_1, ..., X_{95}, C_1, ...C_5$ are our features. $X_1, ..., X_{95}$ are float and $C_1, ...C_5$ are categorical with unique values J, G, H, K and I. Also, the dataset has many missing values.

By examining the categorical variables, we notice that C_2 has only the unique value 'H', thus we remove this variable from our data since the lack of variation means the variable does not contribute any useful information for predicting the target variable.

2 Model and Method

2.1 Model Selection

In the initial phase of our analysis, we considered employing Ordinary Least Squares (OLS). OLS is a classic technique known for its simplicity and interpretability, making it a logical starting point for linear modeling. However, OLS assumes that all predictors are equally important and ignores the possibility of overfitting when dealing with high-dimensional data or when predictors are correlated. Given the complexity of our dataset (100 observations with 100 features), overfitting occurs because OLS will always use all 100 features, potentially leading to a model with high variance that captures noise as if it were signal. This can result in poor predictive performance when the model is applied to new, unseen data.

To handle these issues, we explore regularization techniques that introduce a penalty for larger coefficients, such as Ridge Regression (which imposes an L2 penalty), Lasso Regression (which imposes an L1 penalty), Elastic Net (which is a combination of both L1 and L2 penalties), and Least Angle Regression (LARS). Our target variable y is continuous and our problem is supervised, that is why we consider regression methods . Ridge Regression helps in shrinking the coefficients and is particularly useful when many features are correlated. Lasso Regression, on the other hand, can set some coefficients to zero, thus performing feature selection and potentially improving model interpretability. Elastic Net combines the properties of both Lasso and Ridge, aiming to maintain grouping effects and variable selection. Lastly, LARS is efficient for high-dimensional data and can be used with any of the previously mentioned regularization techniques. By employing these methods, we aim to achieve better generalization on new data, reduce overfitting, and improve overall model

performance in the context of a "p approximately equal to n" scenario.

2.2 Missing Values

In order to handle the missing values and ensure the integrity of our predictive model, given the potential variations in the underlying reasons for missingness, we used a function that considers two common scenarios: Missing Completely At Random (MCAR) and Missing At Random (MAR).

This strategy involves iterating through each feature column in the dataset and applying one of two imputation techniques based on a criterion that randomly classifies a column as MCAR or MAR. By distinguishing between MCAR and MAR, we tailor our imputation technique to better reflect the underlying data generation process

For columns considered as MCAR,where missing data occur randomly, without connection to any measured or unmeasured attribute, we used a simple random sampling technique. Specifically, for any missing entry in such a column, we randomly selected a non-missing value from the same column to fill the gap. This approach is predicated on the assumption that the observed distribution of values in a column is a reliable representative for the missing data.

For columns classified as MAR, where the occurrence of missing values is determined by the data we have, rather than the data we don't, we use the method of multiple imputation. We fill missing entries based on multivariate imputation by chained equations (MICE). This method works by taking turns with each piece of missing data. For each one, it guesses the missing value using information from the other features that don't have missing data. It keeps doing this in a cycle, getting better and more accurate with each round. This way, we get a well-thought-out guess for what the missing data could be.

2.3 Factor Handling

Before dealing with the missing values of the dataset, we had to handle the categorical features by using a One-Hot Encoding technique. This method transforms each category value into a new binary column, indicating the presence (1) or absence (0) of each categorical feature. After encoding, the original categorical columns are removed from the dataset and the new binary columns are merged back in.

This process avoids creating a false ordinal relationship among categories and allows our model to interpret and utilize the categorical data effectively, enhancing the accuracy of our predictions.

3 Model Validation

In the model validation section of our report, it is important to discuss the measures taken to ensure the robustness and predictive accuracy of our regression models. This approach aligns with best practices in machine learning, as it facilitates the tuning of model

hyperparameters to enhance generalizability while preventing information leakage from the evaluation set during the training process.

During cross-validation, the data in the larger subset was systematically divided into five folds, with each fold serving once as a validation set while the others constituted the training set. This method allowed us to iteratively train and validate the models across diverse segments of the data, thereby reducing the likelihood of overfitting and providing a comprehensive understanding of model performance. At this point, we, also, used the two functions for the missing values and the handling of the categorical variables on the X_{train}andX_{test} sets, and afterwards we standardized them using for both of them the mean of the train set.

The metrics gathered from the cross-validation R-squared and RMSE were recorded and analyzed for each iteration and across an expansive range of hyperparameters, including different values of lambda log(-3, 3, 150) for Ridge, Lasso, Elastic Net, and LARS regressions. These efforts were crucial for determining the optimal balance between bias and variance, ultimately leading us to the best hyperparameters for each model.

Regression Method	Best λ	Best R2 Score	Best RMSE	Best λ_{Cp}
Ridge	29.498	0.6481	30.4364	273
Lasso	2.413	0.7195	26.3021	12.80
Elastic Net	0.031	0.6723	28.8962	0.078
Lars	19	0.6927	28.2253	8

Table 1: Comparison of Regression Model Performances

We examine the R^2 score and RMSE to evaluate our models comprehensively: R^2 offers insight into how well our model explains the variability of the data, while RMSE measures the accuracy of our predictions by quantifying the average difference between the observed and predicted values. Together, they provide a balanced assessment of model performance.

Furthermore, we consider the Cp statistic to measure the model efficiency, balancing between model complexity and fit. A lower Cp suggests a model with fewer predictors without compromising on predictive power. Comparatively, models closest to the ideal Cp value, where Cp $\approx$ number of predictors, indicate an optimal trade-off.

- Ridge Regression exhibited a moderate performance with an R^2 score of 0.6481 and an RMSE of 30.4364. The optimal lambda for Ridge was determined to be 29.498, indicating a relatively high level of regularization to prevent overfitting. The fact that the $\lambda_{Cp} = 273$ that minimizes the Cp curve is so much larger than the λ that minimizes the RMSE indicates that the Ridge model overfit.

- Lasso Regression emerged as the standout model, achieving the highest R^2 score among the contenders at 0.7195 and the lowest RMSE of 26.3021. This suggests that Lasso was the most effective at both explaining the variance and accurately predicting the outcomes, benefiting from an optimal lambda of 2.413. Also, we can see that λ_{Cp} is close to λ. The ability of Lasso to perform variable selection likely contributed to its superior performance by reducing model complexity and focusing on the most informative features.

- Elastic Net Regression presented a balanced performance with an R^2 score of 0.6723 and an RMSE of 28.8962. Its optimal lambda was 0.031, indicating a nuanced compination of Ridge and Lasso's regularization techniques. Elastic Net's ability to combine these approaches offered a compromise between feature shrinkage and selection.

- Lars Regression showed a competitive performance with an R^2 of 0.6927 and an RMSE of 28.2253, with an optimal lambda value of 19. This method's performance underscores its efficiency in handling high-dimensional data, striking a balance between complexity and predictive accuracy.

Considering both the R^2 score and RMSE as measures of model success, Lasso Regression is identified as the best performing model for our dataset. It not only provided the highest explanation of variance but also demonstrated the greatest accuracy in predictions, as reflected by the lowest RMSE.

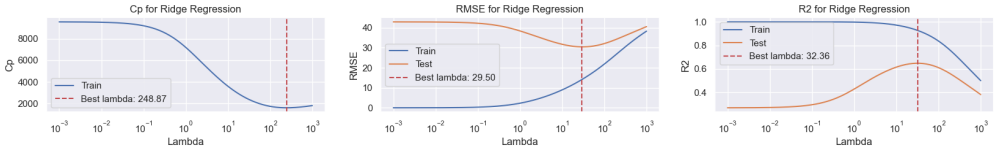

Figure 1: Ridge Regression

Figure 2: Lasso Regression

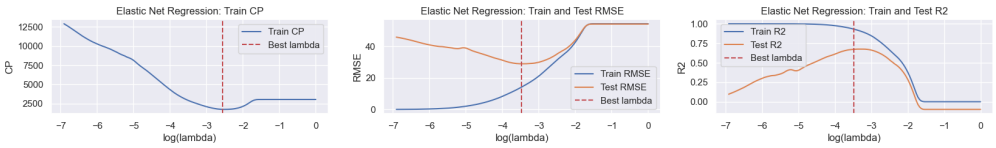

Figure 3: Elastic Net Regression with l1 ratio = 0.9

Figure 4: Least Angle Regression

By examining the Cp-statistic, we conlcude that Ridge Regression's higher optimal lambda value suggests it's less ideal for simplicity, despite handling multicollinearity well. Lasso Regression strikes a better balance between simplicity and accuracy, as indicated by its lower optimal lambda and Cp minimum, suggesting a simpler, interpretable model through feature selection. Lars Regression's moderate lambda choice reflects a fair balance, possibly fitting for certain prediction needs. Elastic Net's Cp, higher than Lasso's but still reasonable, could suit data needing a mixed regularization approach. Overall, Lasso stands out for its optimal mix of simplicity and predictive strength.

Thus, after taking into considaration R^2, RMSE and Cp-statistic, we choose as our best approach Lasso Regression and we have two values for best lambda.

3.1 Bootstrap

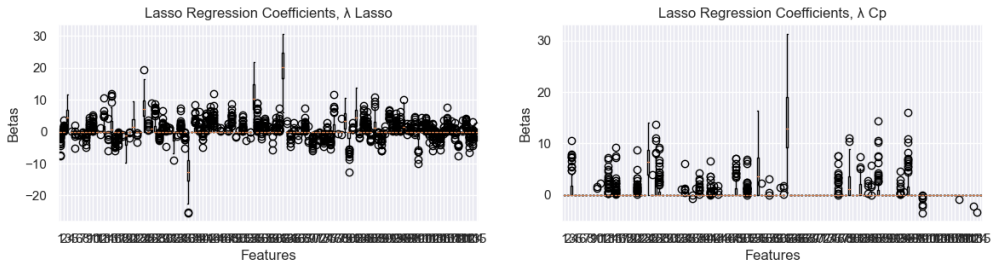

Figure 5: Bootstrapping

The plot on the left indicates a greater number of features with coefficients shrunk towards zero, suggesting a simpler model, while the plot on the right shows less shrinkage and potentially a richer model with more features contributing. Considering the trade-off between model complexity, interpretability, and predictive power, the model with λ selected via minimizing the RMSE (left) would generally be preferred for prediction, as it suggests a model that balances complexity with robustness to overfitting.

4 Results

In the final stage of our analysis, we focus on a new dataset, consisting of 1000 additional observations. Maintaining consistency in our approach, we implemented the same preprocessing steps that were applied to the initial training data. This ensures that our predictive model, a Lasso Regression fine-tuned with a lambda of 2.413, processes the new data under the same conditions that provided optimal results previously.

For the calculation of the estimation of the prediction error $\hat{\text{RMSE}}$ on the new dataset, we adopted the mean RMSE from cross-validation. This choice is grounded in the assumption that the new dataset closely mirrors the distribution and characteristics of the data used during the cross-validation process. Cross-validation, by evaluating the model across multiple subsets, provides a robust and stable estimate of the model's generalization error.

List of Figures

List of Tables