Assignment 2 - Ames House Price Dataset

Hengjian (Henry) Jia

December 1, 2021

1 Dataset

The dataset I've chosen is the Ames House Price dataset (`http://jse.amstat.org/v19n3/decock.pdf`). This dataset is widely used in regression analysis tutorials, so is suitable for this assignment. It is collected by Dean De Cock from the Truman State University. The dataset contains the value of individual residential property in Ames, Iowa from 2006 to 2010. The dataset contains 2930 observations and a number of explanatory variables.

Predicting house prices is obviously useful for a number of reasons. Estate agents can use this to accurately gauge prices for negotiating sales. Economists can use it to help gauge economic metrics. But, I've only got 2 pages so I can't rabbit on.

2 Analysis

The vanilla linear regression technique introduced in the lectures are absolutely hopeless here. The large number of variables caused the model to overfit wildly. Vanilla linear regression yields a root mean squared error (RMSE) of around \$19000, but a test set error of well over several million. There are multiple ways to remedy this. One method is to reduce the number of explanatory variables used in the regression model. The explanatory variables which are used can be chosen by individually evaluating their correlation with the sale price, and then selecting only the variables with the highest correlation to use in the regression model. This can be visualised through plotting the number of highest correlation explanatory variables used, against the resulting RMSE. As shown in figure 1, by selecting the 50 highest correlated variables with the sale price, a RMSE of just over \$35000 can be reached on the test set.

An alternative would be to use ridge regression, a modified version of the vanilla linear regression. This however, requires a regularisation coefficient λ to be chosen. This can chosen visually by plotting the value of λ against the resulting RMSE shown in figure 2. This shows that an RMSE of just over \$35000 can be reached when $2^2 \leq \lambda \leq 2^7$.
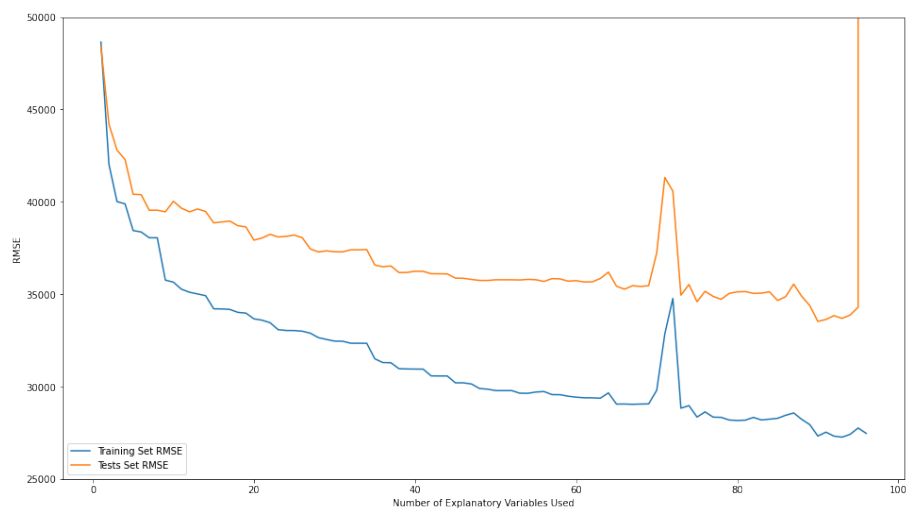
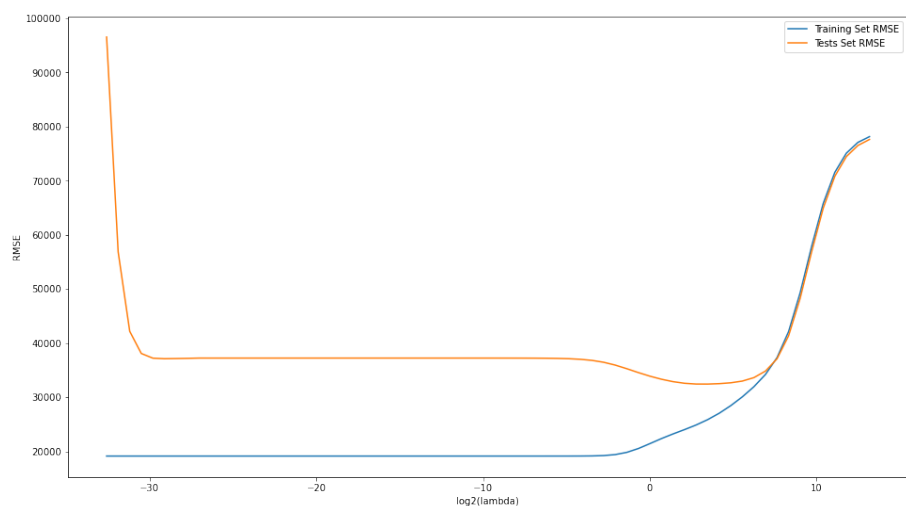Figure 1: Root mean squared error against number of explanatory variables for linear regression

Figure 2: Root mean squared error against $\log_2(\lambda)$ for ridge regression

3   Conclusion

From this analysis, it can be concluded that house prices can be reasonably accurately gauged through adapting the regression techniques in lectures. By choosing the explanatory variables intelligently or using advanced regression techniques, a suitable model can be inferred to predict house prices in Ames, Iowa with an RMSE of just over $35000. The code for this report can be found at `https://github.com/HenryJia/MA930/tree/main/Assignment2`.

A Appendix

A.1 Data Preprocessing

The data needed a cleaning and preprocessing before any analysis. I first purged any explanatory variables which had more than 80% missing data. This lead to 4 variables being purged: the type of road leading to the property, the quality of swimming pool the property had (if it had one), the quality of the property's fence (if it had one), and miscellaneous features of the property. This was followed by filling in any remaining missing entries with the mean for numerical variables, and the mode for categorical variables.

Finally, all numerical variables were shifted and scaled to have mean 0 and variance 1. All categorical variables were converted to a one hot representation.

A.2 Ridge Regression

Ridge regression seeks to minimise the modified version of the mean squared error

$$E(W) = \frac{1}{N} \sum_{k=1}^{N} (b + W^T x_k - y_k)^2 + \lambda ||W||_2^2$$

Where λ is a hyperparameter. λ controls how much we penalise the model for making W large. This incentivises the model to not fully utilise all the input variables and reduces overfitting. The Bayesian interpretation of this is this is equivalent to applying a 0 mean Gaussian prior on W. By adjusting λ we can balance overfitting and underfitting. If λ is set to 0, this is equivalent to vanilla linear regression. If λ is large, then the model tends to towards a horizontal straight line.