International Institute of Information Technology Bangalore

upGrad

Advanced Regression Assignment

Author: *Anish Mahapatra*
Machine Learning II > Module 3

Q1. What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

A: The optimal values for alpha in ridge and lasso regression are as follows:

- Ridge Regression: Optimal value of alpha is 50
- Lasso Regression: Optimal value of alpha is 0.03

If we were to double the value of alpha in Ridge Regression, then the Negative Mean Absolute error of the Train Score decreases and the Negative Mean Absolute error increases. The bias of the model decreases and there may be an increase of variance of the model.

If we were to double the value of alpha in Lasso Regression, then the R^2 value of the *train* dataset reduces from around 84% to 77% and the value f R^2 score reduces from 82% to 77%. This implies that there is a reduction in the bias and variance of the model, hence, it would not be optimal.

The important predictor variables after we double the value of alpha is as follows:

1. OverallQual: Rates the overall material and finish of the house
2. FireplaceQu_NoFireplace: Indicative of no fireplace existing in the house
3. GrLivArea: Above grade (ground) living area square feet
4. GarageCars : Size of garage in car capacity
5. YearBuilt: Original Construction Date
6. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

Q2. You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

A: Based on the R^2 score, we shall choose to apply the optimal value of alpha on **Lasso Regression**, as the R^2 scores are closer on the train and test data set ensuring that the bias-variance tradeoff is balanced.

After analyzing the performance of Lasso and Ridge Regression, we conclude that Lasso regression has an edge over the Ridge Regression model as the train and test scores are comparatively higher. More iterations can be performed to improve model performance. Cross-validation with 10 folds is also incorporated within the model-building.

The reason Lasso Regression was chosen was it also aided the model in feature selection along with weights. So, it enabled the analyst to inform the business from the approximate 250 feature model, which features to focus on as mentioned below.

The important features from the analysis are as follows:

1. OverallQual: Rates the overall material and finish of the house
2. GrLivArea: Above grade (ground) living area square feet
3. GarageCars : Size of garage in car capacity
4. YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)
5. TotalBsmtSF: Total square feet of basement area
6. YearBuilt: Original construction date
7. PoolQC_Gd: Pool quality

Q3. After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

A: If the top 5 important features are excluded, then the following features shall be chosen:

1. YearBuilt: Original construction date
2. PoolQC_Gd: Pool quality
3. MSZoning_RM: Identifies the general zoning classification of the sale
4. CentralAir: Central air conditioning present
5. FireplaceQu_NoFireplace: Fireplace quality – fireplace not included

Q4. How can you make sure that a model is robust and generalizable? What are the implications of the same for the accuracy of the model and why?

A: Robustness is the property that tested on a training sample and on a similar testing sample, the performance is close. We can ensure that a model is robust if it has a similar cost function around the training data set.

The complexity of the model performing the defined task of regression or classification and the algorithm's generalizability are related. The Vapnik-Chervonenkis (VC) as explained in the course theory provides a general measure of complexity and proves that bounds on errors is function of complexity.

Although reliability is defined in terms of consistency or generalizability, specific statistical indices of reliability will vary depending on the statistical model and the sources of error for the model. Sometimes, even if there are more misclassified training points but the model can generalize better, it is to advisable to apply models to future data rather than achieve the best fit to existing data.

- Sometimes, if we increase the robustness or generalizability of the model, there tends to be a **decrease in the accuracy** of the model, especially on the training dataset. However, as models made are to be applied on unseen testing datasets, it is better to have a more generalizable and robust model by having a relatively simple model and using only as many features as necessary respectively.