Advanced Regression Assignment

Student: Karthik Jayaraman

Assignment-based Subjective Questions

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

   (Answer)

   a. Optimal value of alpha for ridge regression is 0.2
   and Optimal value of alpha for lasso regression is 0.0001.

   b. The changes in the model if you choose to double the value of alpha for both ridge and lasso, the following is observed:
   For Ridge regression:
      i. The R2 score of Train and Test a slight change but almost same
      ii. The Mean Squared error of Train has slight increase.

	Metric	Ridge Alpha=2.0	Ridge Alpha=4.0
0	R2 Score (Train)	0.944150	0.939895
1	R2 Score (Test)	0.905723	0.906991
2	RSS (Train)	0.738517	0.794774
3	RSS (Test)	0.492228	0.485609
4	MSE (Train)	0.029019	0.030104
5	MSE (Test)	0.036182	0.035938
6	No Of variables	183.000000	183.000000

For Lasso regression:

i. The R2 score of Train and Test slightly decreased.
ii. The Mean Squared error of Train & Test has increased.

	Metric	Ridge Alpha=2.0	Ridge Alpha=4.0
0	R2 Score (Train)	0.940216	0.934196
1	R2 Score (Test)	0.906370	0.904946
2	RSS (Train)	0.790535	0.870143
3	RSS (Test)	0.488851	0.496286
4	MSE (Train)	0.030023	0.031499
5	MSE (Test)	0.036057	0.036331
6	No Of variables	103.000000	103.000000

c. The most important predictor variables after change is implemented:

Ridge Regression:

i. GrLivArea: Above grade (ground) living area square feet
ii. OverallQual: Rates the overall material and finish of the house
iii. TotalAreaSF: (Derived variable) Total Area in square feet
iv. BsmtFinSF1: Type 1 finished square feet
v. OverallCond: Rates the overall condition of the house
vi. Neighborhood_Crawfor: Physical locations within Ames city limits (Crawford)
vii. BsmtQual_Gd: Good (90-99 inches) height of the basement
viii. BsmtUnfSF: Unfinished square feet of basement area
ix. BsmtQual_TA: Typical height (80-89 inches) of the basement
x. KitchenQual_TA: Kitchen quality (Typical/Average)

Lasso Regression:

a. TotalAreaSF: (Derived variable) Total Area in square feet
b. OverallQual: Rates the overall material and finish of the house
c. GrLivArea: Above grade (ground) living area square feet
d. OverallCond: Rates the overall condition of the house
e. YearBuilt: Original construction date
f. BsmtFinSF1: Type 1 finished square feet
g. Neighborhood_Crawfor: Physical locations within Ames city limits (Crawford)
h. BsmtQual_Gd: Good (90-99 inches) height of the basement
i. LotArea: Lot size in square feet
j. BsmtExposure_Gd: Good Exposure to walkout or garden level walls

2. 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?
(Answer)
   a. The Optimal value of alpha for ridge regression is 0.2
   and Optimal value of alpha for lasso regression is 0.0001.

   b. The R-squared value (R2 score) is slightly higher in Lasso regression for Test.
   c. The Mean squared Error (MSE) is slightly lower in Lasso regression for Test.
   d. The number of variables or features in Lasso regression (103) is lesser than Ridge Regression (183). This could be because Lasso shrinks coefficients of insignificant features to exactly zero.
   e. Lasso Regression model is simpler than Ridge regression model.
   f. Lasso Regression model increases model interpretation because of reduction in features.
   g. Lasso regression model could have improved computing efficiency.

   Thus, I will choose Lasso regression model to apply in assignment.

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.940077	0.944150	0.940216
1	R2 Score (Test)	0.882806	0.905723	0.906370
2	RSS (Train)	0.792366	0.738517	0.790535
3	RSS (Test)	0.611879	0.492228	0.488851
4	MSE (Train)	0.030058	0.029019	0.030023
5	MSE (Test)	0.040340	0.036182	0.036057
6	No Of variables	94.000000	183.000000	103.000000

3. After building the model, you realised 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? **(Answer)**

The Top 5 most important predictor variables in the lasso model are dropped and a new lasso model is created.
The Five most important predictor variables are:

a. BsmtFinSF1: Type 1 finished square feet
b. BsmtUnfSF: Unfinished square feet of basement area
c. BsmtFinSF2: Type 2 finished square feet
d. TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
e. ExterQual_Fa: Fair quality of the material on the exterior

4. How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?
(Answer)

A machine learning model is created by fitting a statistical function to training data set by tuning the hyperparameters.

Let us take an example:
If manufacturer builds a bike and if its suspension is too tight or tire pressure is too high, the ride might not be comfortable. Its overfitted to the ideal suspension metrics and tire pressure.

If manufacturer builds a bike and if its suspension is too loose or tire pressure is too low, the ride might not be comfortable again. Its underfitted and far from the ideal suspension metrics and tire pressure.

If the suspension setting and tire pressure is just right, the ride would do well in gravel, smooth or bad road conditions and the ride would be comfortable.

A machine learning model (like a bike) is similar where it has to flexible (generalisable) and robust enough to adapt to various real world (test data) conditions (Not underfitting or overfitting to training data. Just right).

Regularization techniques helps to manage model complexity by shrinking the coefficient estimates towards zero and avoid risk of overfitting. This is done by adding a penalty to the cost term.

Cost = Residual Sum of Squares (RSS + Penalty)

If Bias (too simplistic) in a model is HIGH, when it does not perform well on training data.
If Variance (too sensitive) is HIGH, when model does not perform well on test data.

Bias and variance should be both low (in a balance). Regularization helps to manage this.

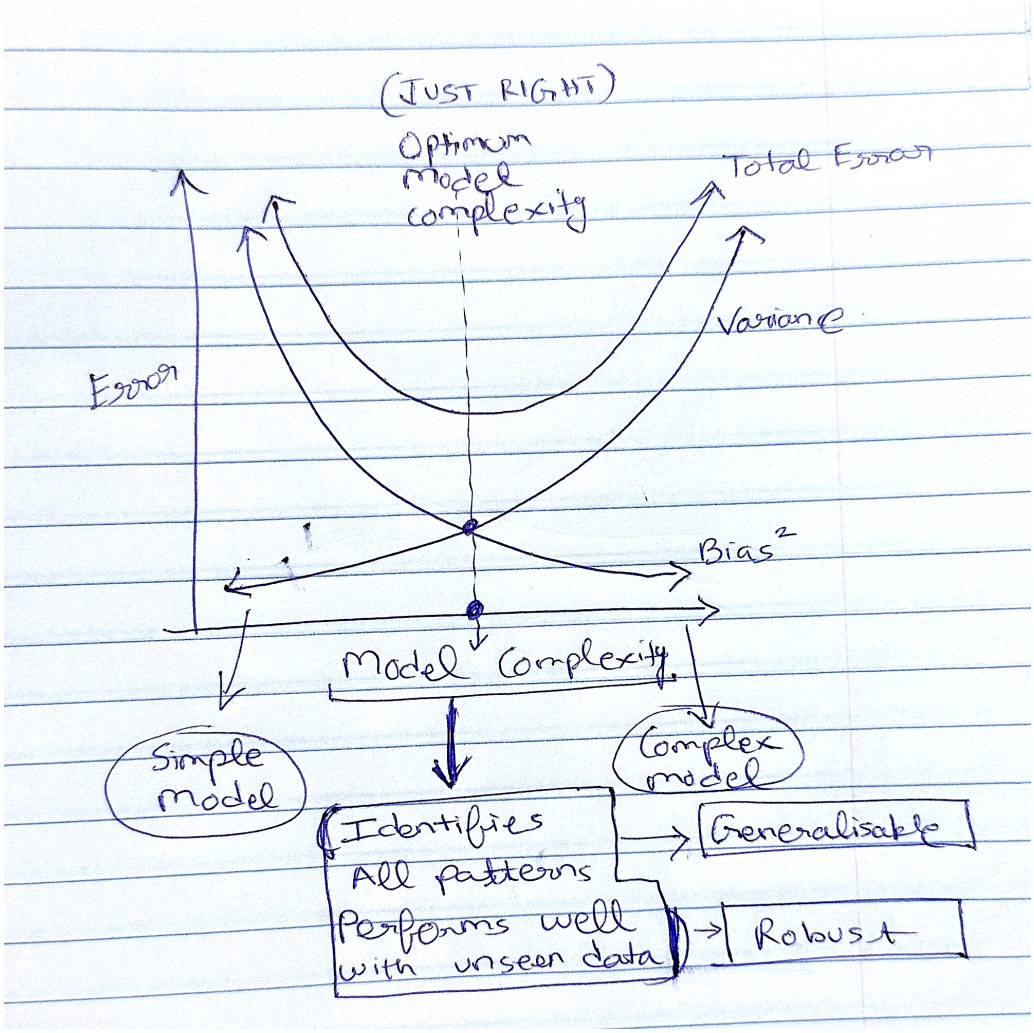


Figure 1: Bias-Variance Tradeoff and Model complexity

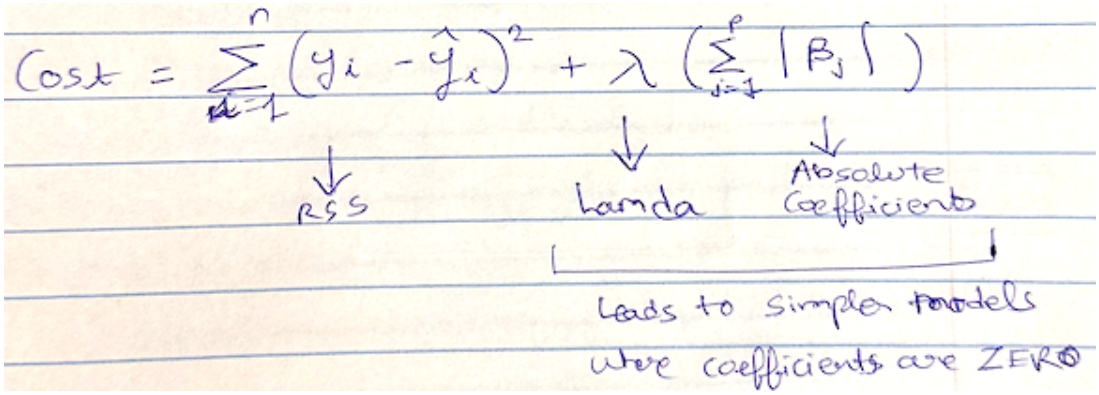
Ridge regression Cost:

$$Cost = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

RSS — Lambda — Coefficients

Regularization Term

Lasso regression Cost:



Both Ridge and Lasso regression shrinks the coefficients towards zero.
But Penalty in Lasso forces coefficient estimates to be exactly zero thereby eliminating some features. This may result in lower accuracy of model.

Regularization increases bias and decreases variance (avoids complexity and overfit).