upGrad

#LifeKoKaroLift

Advanced Regression: Assignment

Course : Data Science

Session : Advanced Regression Pre-Assignment

Instructor : Dr Reena Duggal

upGrad

What we will cover in this session?

1 The problem of Overfitting, Underfitting and Regularisation

2 Trade off between error term and regularisation term

3 Assignment walkthrough

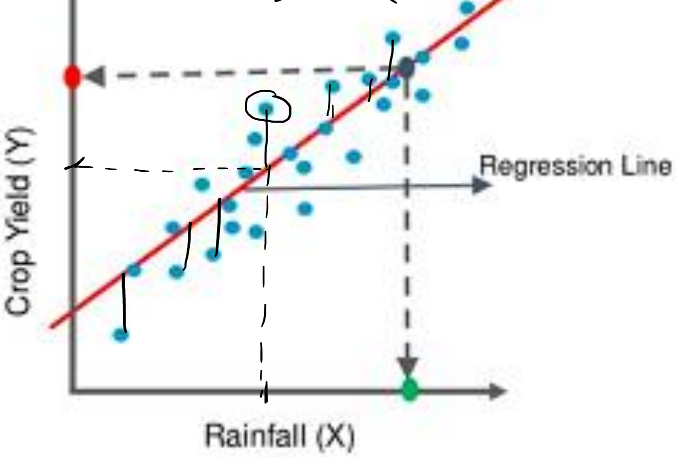
4 QnA

Linear Regression quick recap

OLS

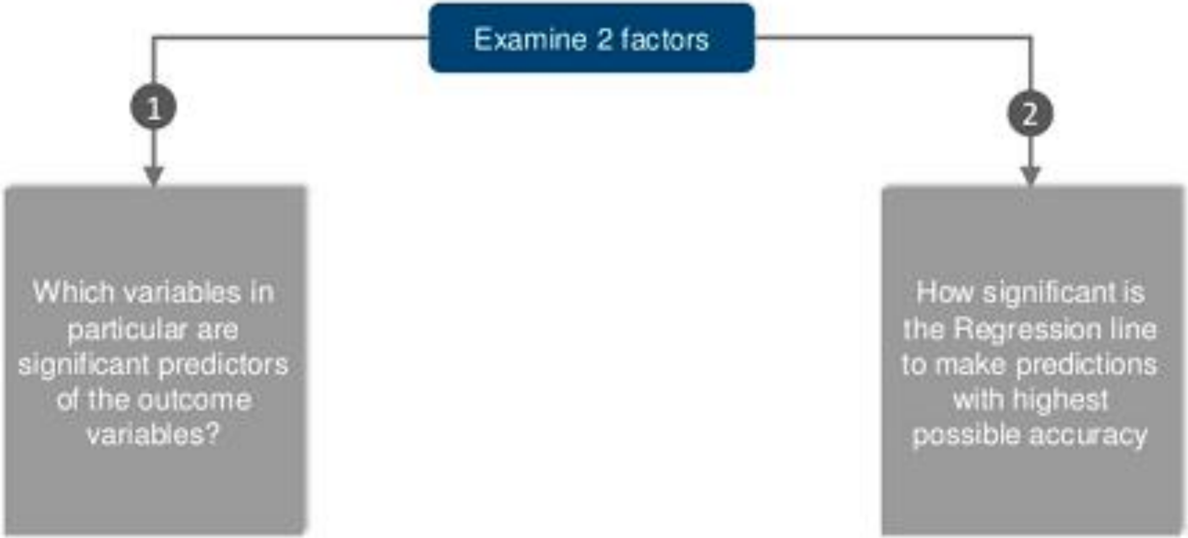
Ordinary Least Square Method

Cost Function = $\Sigma(\text{Error})^2 = \Sigma(y - \hat{y})^2$



$\beta_0, \beta_1, \beta_2$ - - - - - -

Linear Regression is a statistical model used to predict the relationship between independent and dependent variables.



Examine 2 factors

1. Which variables in particular are significant predictors of the outcome variables?
2. How significant is the Regression line to make predictions with highest possible accuracy

Linear Regression: Single Variable

$$\underbrace{\boxed{\hat{y}}}_{\text{Predicted output}} = \underbrace{\beta_0 + \beta_1}_{\text{Coefficients}} \underbrace{\boxed{x}}_{\text{Input}} + \underbrace{\boxed{\epsilon}}_{\text{Error}}$$

Linear Regression: Multiple Variables

$$\boxed{\hat{y}} = \underbrace{\beta_0 + \beta_1}_{} \boxed{x_1} + \dots + \underbrace{\beta_p}_{} \boxed{x_p} + \boxed{\epsilon}$$

Overfitting and Underfitting

Total Error = Bias + Variance

What is Bias and Variance

$y = \underline{100}x_1 + \underline{200}x_2$

High var.

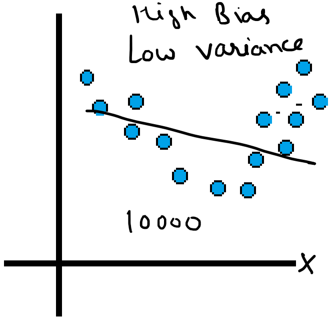
$y = 10,000$

$y = 12,000$

not generalized

High Bias
Low Variance

Not Learning the data or pattern
"Underfitting"

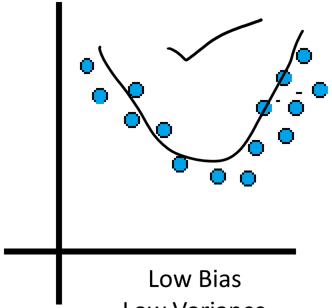


Features from Raw Data
Area (x_1)
# of Bedrooms (x_2)
Size of Bedrooms = $\frac{\text{Area}}{\text{\# of Beds}} = \frac{x_1}{x_2}$

$y = \beta_0 + \beta_1 x$

Degree, Features ⇒ Fixed

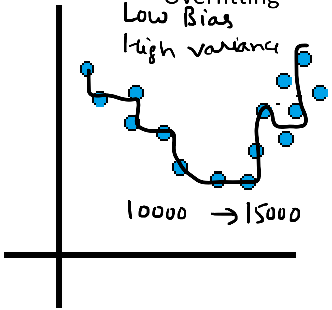
Learning the pattern
and not the data
"Good model"



Low Bias
Low Variance

$y = \beta_0 + \beta_1 x + \beta_2 x^2$

Learning the data and not
Learning the pattern
"Overfitting"



Low Bias
High Variance

$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + ----$

$y = \underline{\beta_0'} + \underline{\beta_1'} x_1 + \underline{\beta_2'} x_2^2$

Complex = Big

Model is rigid, not generalized

Bias - Amount of Error that the model is making

Variance - Amount that the model(target variable) will change given different training data

One approach to reduce the variance in the model is to shrink the coefficient estimates towards zero.

How to solve Overfitting?

Information loss
↓

Simplify the model

RFE

1. Reduce the number of features
   a. Select the best features based on business need
   b. Use some model selection algorithm

p-value (<0.05, predictor is significant
≥ 0.05, " " not " ")

VIF (Multicollinearity)

But what works well is what is known as "**Regularisation**"

2. Regularisation
   a. Regularization adds a penalty on the different parameters of the model to reduce the freedom of the model.

   $\beta_1, \beta_2, \beta_3, +$ ----

   b. Hence, the model will be less likely to fit the noise of the training data and will improve the generalization abilities of the model
   c. It will keep all the features but reduce their value of coefficient on the target variable.
   d. We keep all the variables and didn't lose information.

How to solve Overfitting?

upGrad

Regularisation

λ is in theory

Python sklearn

alpha (lineare Regsn)

$C = \frac{1}{\lambda}$ (logistic Regression)

What is the cost function for linear regression?

Σ (Actual − Predicted)

$$\sum_{i=1}^{n}(x_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2$$

$+ \lambda \left(|\beta_1| + |\beta_2| + |\beta_3| + ----\right) \longrightarrow$ Lasso

How to limit the values of coefficients so that they remain small and doesn't become too complex?

- Add **regularization term** to the **error term**

$$\text{Min}\left[\Sigma\left(y - (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ---)\right) + \underline{\lambda}\left(\beta_1^2 + \beta_2^2 + \beta_3^2 + -----\right)\right] \longrightarrow \text{Ridge}$$

Large

$$\text{Min}\left(\left[10 + 50x_1 + 60x_2 + ---\right] + \left(1000\left(50^2 + 60^2\right)\right)\right)$$

5 6 Focus here 5 6

Focus shifts here

$50x_1 + 60x_2$

$+ 1 \left(50^2 + 60^2\right)$

How to solve Overfitting?

upGrad

a) $y = 10x_1 + 20x_2$

b) $y = 100x_1 + 200x_2$

Regularisation

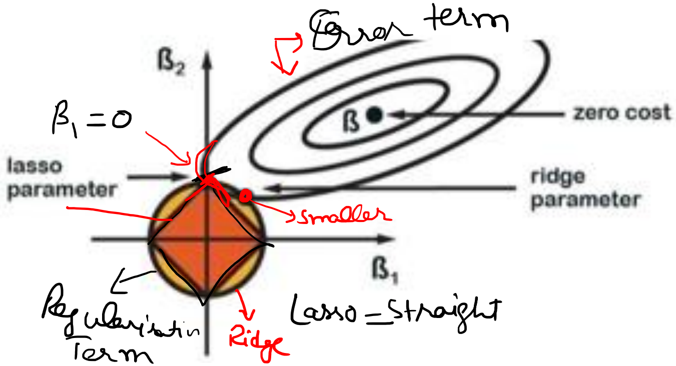
Let’s write the generalised error equation for Regularised Regression.

LASSO

$$\underbrace{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2}_{} + \lambda\sum_{j=1}^{p}|\beta_j|$$

RIDGE

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$



Lasso regression not only helps in reducing over-fitting but it can help us in feature selection

Let’s see how the regularisation parameter will affect the fit of the model.

1. λ is high, more regularisation. Model will be simple, but you run the risk of underfitting your data. Your model won't learn enough about the training data to make useful predictions.

2. λ is low, less regularisation. Model will be more complex, and you run the risk of overfitting your data. Your model will learn too much about the particularities of the training data, and won't be able to generalize to new data.

Assignment

upGrad

How house price is related to all the exploratory variables?

Problem Statement

Data Dict

1500 rows

≃ 80 columns

A US-based housing company named Surprise Housing has decided to enter the Australian market. The company uses data analytics to purchase houses at a price below their actual values and flip them at a higher price. For the same purpose, the company has collected a data set from the sale of houses in Australia.

The company is looking at prospective properties to buy to enter the market.

Assignment

upGrad

What you need to do?

- Which variables are significant in predicting the price of a house?
- How well those variables describe the price of a house? Accuracy
- Determine the optimal value of lambda for ridge and lasso regression.

Data walkthrough.

Assignment

upGrad

Steps to proceed with the assignment:

Data Cleaning
Data types
Univariate
Bivariate

1. Perform EDA to understand the data
2. Check missing values
   - Actually missing   # of Bedroom
   - Missing with some meaning

     Garage Quality (Good, Bad) NAN → None
     Fence → None
     Alley
     Basement Quality
   - Drop columns with high percentage of missing values >80% → Drop
   - For continuous variables, try to impute the missing values with mean or median. Perform EDA to find out which one fits best

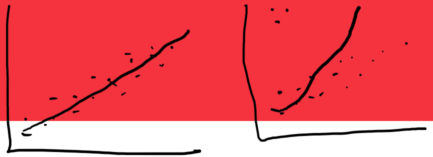
     median ↓ outliers
   - For categorical columns, try to impute with mode

Basement Quality < 90% None
10% (Good, Bad)

↓ Information Content is less

Value-counts

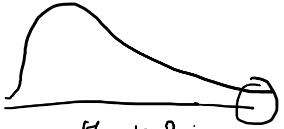
Assignment

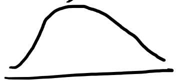


x	log
1	0
10	1
100	2
1000	3

Steps to proceed with the assignment:

3. Check Outliers Data Entry Problem ⇒ Missing values
   - Artificial Outliers
   - Natural Outliers → Transformation (Log, Sqrt, Cuberoot, Minmax, Standardization)
     - Check if the target variable is normally distributed or not? Log(House Price)





4. Create dummies for categorical data pd.get-dummies( )
   - ≈200 You can create groups of the categories to reduce the number of categories and then create dummies. This is an optional method.
5. Handling year columns
   - There are 4 columns that contain year. What to do with them?
   - How to convert them?

Year Built
Year Sold
Year Remodeld
Garage Build

Age of the House
(Year Sold - Year Built)

How to reduce dummy variables?

① Assign Class group

Category	Cat group
x	food
y	food
z	drink
a	drink
b	Electronic

Need to have domain knowledge

② Combine different small categories

Category	
x	40%
y	40%
z	5%
a	5%
b	10%

x	40%
y	40%
Others	20%

How to solve..Continue.

Split (70%, 30%)
Scaling
Divide X, Y

Params = [0, 0.2, 0.5, 0.1, - - -
- - -, 20, 500, 1000]

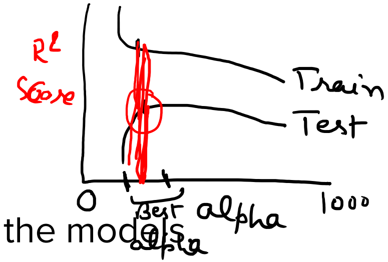
gridsearchcv (sklearn. lasso()
ridge())

What to do?

Feature Elimination < RFE
VIF

Model Building:

1. You can directly start trying out Lasso and Ridge with different values of alpha
   - For Lasso choose the best alpha
   - For Ridge choose the best alpha

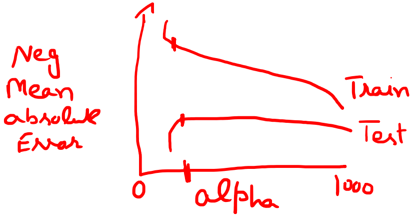


After choosing the best alpha from both the models, check the performance of both the models.

Lastly, you need to find out best features that describes the price of the house, for this check the top features for both final models created using Ridge and Lasso and then choose the features accordingly.

$$y = 0.01 + \boxed{0.5} \times \underline{\text{Area}} + \boxed{1.2} \times \text{Parking} + 0.03 \times \text{Basement} + \text{- - - - - -}$$

Top 2 → Most Coeff.



How to solve..Continue.

What to do?

Subjective Questions: 4-5

You need to answer these question using your learnings from the module Advanced Regression and the model you have obtained.

① Ridge alpha = 0.5 x 2 = 1.0 Train & Test Accu

Lasso alpha = 0.1 x 2 = 0.2

② Lasso Top 5 next Top 5 Predictors

~~Area~~

~~Pool~~

~~Parkin~~

~~Garage Quality Good~~

Pavement

What to keep in mind

Swimming-Pool-avo

10000 90000

None

yes
No
yes
No
10%

Yes
Yes
No
No

20%

90%

- Add comments after every cell of code. So that we can understand your approach and method.
- Describe the results.
- For subjective answers, use DOC and type on it, if you wish to add images you can. But convert it to PDF before submitting.
- Create only one Jupyter notebook.
- Submit one zip file with the code and the PDF.

Params = [0.1, 0.5, 0.7, 0.9, 1]

GridSearchCV [lasso, alpha params = Params]

1st alpha = 0.1

2nd 0.5

Ordinal

Basement Quality

3 Excellent

2 Good

1 Bad

0 None

↑

Price

↑

Nominal

Quiz Time

Poll Questions

Question-1: Suppose we have a regularized linear regression model: $\text{argmin}||Y-\beta x||^2+\lambda||\beta||$. What is the effect of increasing λ too much on bias and variance?

a) Increases bias, increases variance
b) Increases bias, decreases variance
c) Decreases bias, increases variance
d) Decreases bias, decreases variance
e) Not enough information to tell

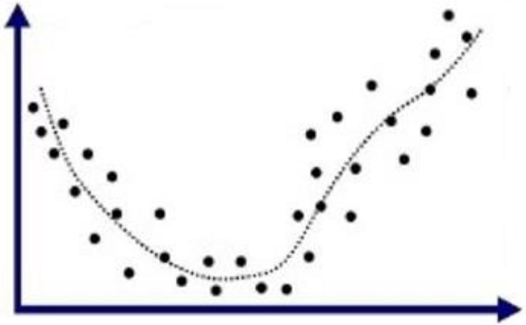
Question-2: After applying a regularization penalty in linear regression you find that some of the coefficients are zeroed out. Which of the following penalties might have been used?

a) L0 norm
b) L1 norm
c) L2 norm
d) either (A) or (B)
e) any of the above

Quiz Time

upGrad

Poll Questions



Question-3: Suppose you used a degree-3 polynomial to fit the data and it look like as shown in the image, what will happen if we use a degree-4 polynomial?

a) There are high chances that degree 4 polynomial will over fit the data
b) There are high chances that degree 4 polynomial will under fit the data
c) Can’t say
d) None of these

Question-4: Which of the following is true for the given fit(refer to the image)?

a) Bias will be high, variance will be high
b) Bias will be low, variance will be high
c) Bias will be high, variance will be low
d) Bias will be low, variance will be low

Quiz Time

upGrad

Poll Questions

Question-5: Suppose, you got a situation where you find that your linear regression model is under fitting the data. In such situation which of the following options would you consider?

1. I will add more variables
2. I will start introducing polynomial degree variables
3. I will remove some variables

a) 1 and 2
b) 2 and 3
c) 1 and 3
d) 1, 2 and 3

Question-6: If the fitted model is underfitting then which of following regularization algorithm would you prefer?

a) L1
b) L2
c) Any
d) None of these

upGrad

Thank You!