

# **MACHINE LEARNING**

1. Which of the following methods do we use to find the best fit line for data in linear regression?

Ans. Least Square Error

2. Which of the following statement is true about outliers in linear regression?

Ans. Linear regression is sensitive to outliers

3. A line falls from left to right if a slope is ?

Ans. Negative

4. Which of the following will have symmetric relation between dependent variable and independent variable?

Ans. Regression and Correlation

5. Which of the following is the reason for over fitting condition?

Ans. High Variance and Low bias.

6. If output involves label than that model is called as:

Ans. All of the above

7. Lasso and Ridge regression techniques belong to

Ans. Regularization

8. To overcome with imbalance dataset which technique can be used?

Ans. SMOTE

9 The AUC Receiver Operator Characteristic( AUCROC) curve is an evaluation metric for Binary classification problems. It uses \_\_\_\_\_ to make graph?

Ans. Recall and precision

10 In AUC Receiver Operator Characteristic (AUCROC) curve for the better model area under The curve should be less

Ans. True

11 Pick the feature extraction from below.

Ans. All of the above

12 Which of the following is true about normal equation used to compute the coefficient of the Linear regression ?

Ans. We don't have to choose the learning rate. TRUE

It becomes slow when number of features is very large. TRUE

### 13 Explain the term regularization?

This is a form of regression that constrains / regularizes or shrinks the coefficient estimates towards zero. In this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting. A simple relation for linear regression looks like this. Regularization is a technique used for tuning the function by adding an additional penalty term in the error function. The additional term controls the excessively fluctuating function such that the coefficients don't take extreme values.

Regularization consists of different techniques and methods used to address the issue of overfitting by reducing the generalization error without affecting the training error much. Choosing overly complex models for the training data points can often lead to overfitting. Regularization is a set of techniques that can prevent overfitting in neural networks and thus improve the accuracy of a Deep Learning model when facing completely new data from the problem domain.

### 14. Which Particular Algorithms are used for regularization?

Ans. *One of the major aspects of training your machine learning model is avoiding overfitting. The model will have a low accuracy if it is overfitting. This happens because your model is trying too hard to capture the noise in your training dataset. By noise we mean the data points that don't really represent the true properties of your data, but random chance. Learning such data points, makes your model more flexible, at the risk of overfitting.*

Regularization

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates towards zero. In other words, *this technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.*

A simple relation for linear regression looks like this. Here  $Y$  represents the learned relation and  $\beta$  represents the coefficient estimates for different variables or predictors( $X$ ).

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

The fitting procedure involves a loss function, known as residual sum of squares or RSS. The coefficients are chosen, such that they minimize this loss function.

$$RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2.$$

Now, this will adjust the coefficients based on your training data. *If there is noise in the training data, then the estimated coefficients won't generalize well to the future data. This is where regularization comes in and shrinks or regularizes these learned estimates towards zero*

## Ridge Regression

$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2 = RSS + \lambda \sum_{j=1}^p \beta_j^2$$

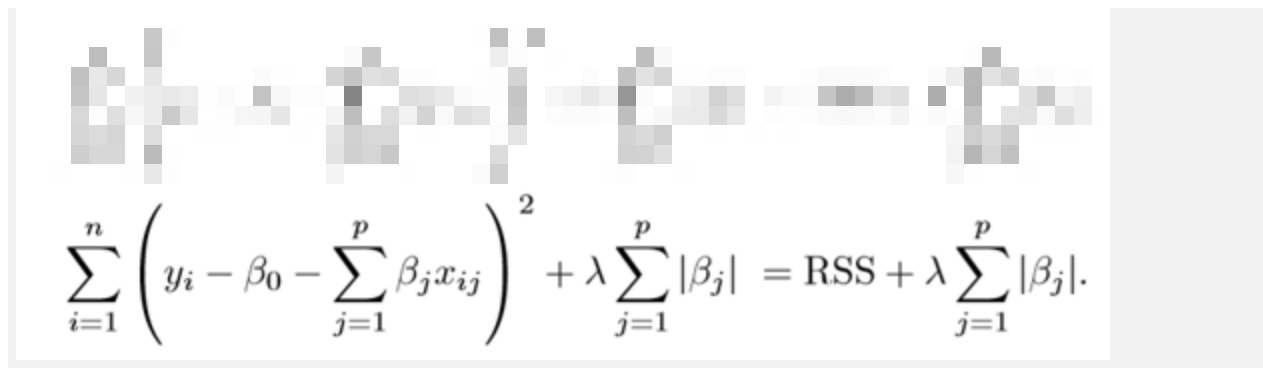
Above image shows ridge regression, where the *RSS is modified by adding the shrinkage quantity*. Now, the coefficients are estimated by minimizing this function. Here,  $\lambda$  is the *tuning parameter that decides how much we want to penalize the flexibility of our model*. The increase in flexibility of a model is represented by increase in its coefficients, and if we want to minimize the above function, then these coefficients need to be small. This is how the Ridge regression technique prevents coefficients from rising too high. Also, notice that we shrink the estimated association of each variable with the response, except the intercept  $\beta_0$ . This intercept is a measure of the mean value of the response when  $x_1 = x_2 = \dots = x_p = 0$ .

When  $\lambda = 0$ , the penalty term has no effect, and the estimates produced by ridge regression will be equal to least squares. However, as  $\lambda \rightarrow \infty$ , the impact of the shrinkage penalty grows, and the ridge regression coefficient estimates will approach zero. As can be seen, selecting a good value of  $\lambda$  is critical. Cross validation comes in handy for this purpose. The coefficient estimates produced by this method are also known as the L2 norm.

The coefficients that are produced by the standard least squares method are scale equivariant, i.e. if we multiply each input by  $c$  then the corresponding coefficients are scaled by a factor of  $1/c$ . Therefore, regardless of how the predictor is scaled, the multiplication of predictor and coefficient ( $x_j \beta_j$ ) remains the same. However, this is not the case with ridge regression, and therefore, we need to standardize the predictors or bring the predictors to the same scale before performing ridge regression. The formula used to do this is given below.

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}}$$

## Lasso


$$\sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| = \text{RSS} + \lambda \sum_{j=1}^p |\beta_j|.$$

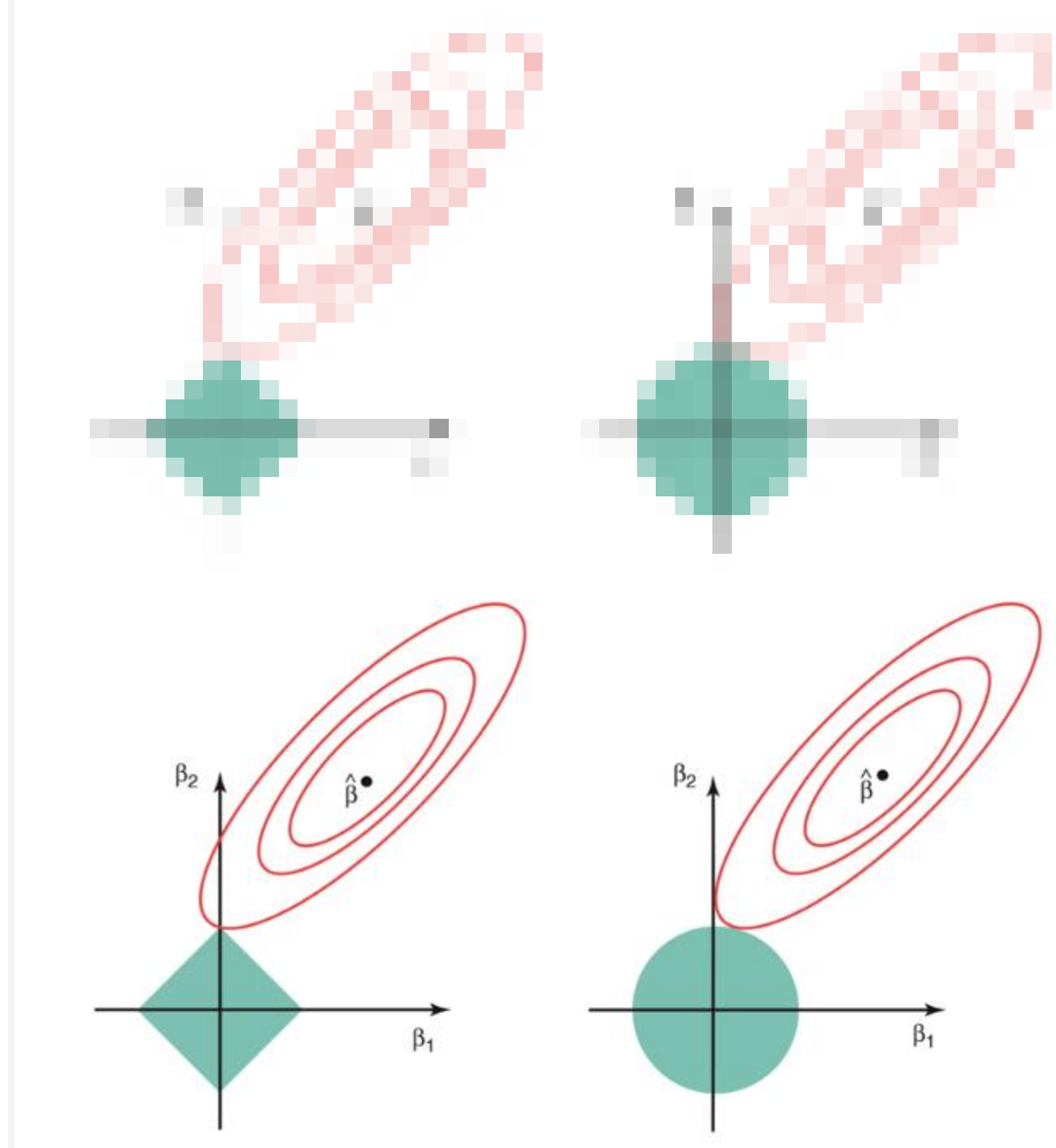
It's clear that this variation differs from ridge Lasso is another variation, in which the above function is regression only in penalizing the high coefficients. It uses  $|\beta_j|$  (modulus) instead of squares of  $\beta$ , as its penalty. In statistics, this is known as the L1 norm.

Let's take a look at above methods with a different perspective. The ridge regression can be thought of as solving an equation, where summation of squares of coefficients is less than or equal to  $s$ . And the Lasso can be thought of as an equation where summation of modulus of coefficients is less than or equal to  $s$ . Here,  $s$  is a constant that exists for each value of shrinkage factor  $\lambda$ . These equations are also referred to as constraint functions.

**Consider there are 2 parameters in a given problem. Then according to above formulation, the ridge regression is expressed by  $\beta_1^2 + \beta_2^2 \leq s$ . This implies that ridge regression coefficients have the smallest RSS(loss function) for all points that lie within the circle given by  $\beta_1^2 + \beta_2^2 \leq s$ .**

**Similarly, for lasso, the equation becomes,  $|\beta_1| + |\beta_2| \leq s$ . This implies that lasso coefficients have the smallest RSS(loss function) for all points that lie within the diamond given by  $|\beta_1| + |\beta_2| \leq s$ .**

The image below describes these equations.



Credit : An Introduction to Statistical Learning by Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani

The above image shows the constraint functions (green areas), for lasso (left) and ridge regression (right), along with contours for RSS (red ellipse). Points on the ellipse share the value of RSS. For a very large value of  $s$ , the green regions will contain the center of the ellipse, making coefficient estimates of both regression techniques, equal to the least squares estimates. But, this is not the case in the above image. In this case, the lasso and ridge regression coefficient estimates are given by the first point at which an ellipse contacts the constraint region. Since ridge regression has a circular constraint with no sharp points, this intersection will not generally occur on an axis, and so the ridge regression coefficient estimates will be exclusively non-zero. However, the lasso constraint has corners at each of the axes, and so the ellipse will often intersect the constraint region at an axis. When this occurs, one of the coefficients will equal zero. In higher dimensions (where parameters are much more than 2), many of the coefficient estimates may equal zero simultaneously.

This sheds light on the obvious disadvantage of ridge regression, which is model interpretability. It will shrink the coefficients for least important predictors, very close to zero. But it will never make them exactly zero. In other words, the final model will include all predictors. However, in the case of the lasso, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter  $\lambda$  is sufficiently large. Therefore, the lasso method also performs variable selection and is said to yield sparse models. A standard least squares model tends to have some variance in it, i.e. this model won't generalize well for a data set different than its training data. *Regularization, significantly reduces the variance of the model, without substantial increase in its bias.* So the tuning parameter  $\lambda$ , used in the regularization techniques described above, controls the impact on bias and variance. As the value of  $\lambda$  rises, it reduces the value of coefficients and thus reducing the variance. *Till a point, this increase in  $\lambda$  is beneficial as it is only reducing the variance (hence avoiding overfitting), without losing any important properties in the data.* But after certain value, the model starts

loosing important properties, giving rise to bias and thus underfitting. Therefore, the value of  $\lambda$  should be carefully selected.

This is all the basic you will need, to get started with Regularization. It is a useful technique that can help in improving the accuracy of your regression models. A popular library for implementing these algorithms is `scikit-learn`. It has a wonderful api that can get your model up and running with **just a few lines of code in python**.

## 15. Explain the term error present in linear regression equation?

Ans. Linear regression most often uses mean-square error (MSE) to calculate the error of the model.

MSE is calculated by:

1. measuring the distance of the observed y-values from the predicted y-values at each value of x;
2. squaring each of these distances;
3. calculating the mean of each of the squared distances.

An error term represents the margin of error within a statistical model; it refers to the sum of the deviations within the regression line, which provides an explanation for the difference between the theoretical value of the model and the actual observed results.

I'm trying to build a regression equation to show the relationship between crop yield (Y) and climatic parameters such as; rainfall (RF), minimum temperature (Tmin) and maximum temperature (Tmax) by using data analysis tool in Excel 2013. There it gives summary statistics including, R-squared, standard error, significance F, interception coefficients. So which term in the summary table can be used as the error term (E) in a regression equation as follows;

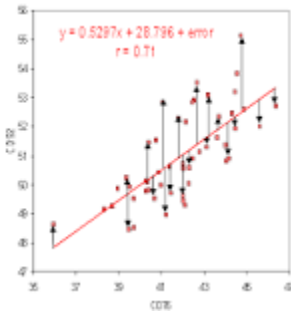
$$Y = a + b \cdot RF + c \cdot Tmax + d \cdot Tmin + E$$

If that is not calculated within the process of excel, how can I calculate and include error term into the equation formed by excel?? (As per my understanding it doesn't provide any error term)

1. The error term is always generated, just by definition, but some programmes, such as Excel, generate it behind the scenes. Excel is certainly not a programme with which you want to do heavy econometrics -- there are more appropriate statistical packages for that.
2. The error term, by definition, is the difference between the actual value of y and its predicted value. The predicted value, again by definition, is  $y = \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \dots + \beta_n \cdot x_n$  for



that concrete observation with concrete values of  $y$  and  $x$ . So if you want to calculate it yourself, you take the actual  $y$  values from the relevant range on your worksheet, calculate the predicted  $y$  for each observation in a separate range and subtract the predicted  $y$  from the actual  $y$ . Voila, you got your error term for all observations. Within a linear regression model tracking a stock's price over time, the error term is **the difference between the expected price at a particular time and the price that was actually observed**. ... The error term stands for any influence being exerted on the price variable, such as changes in market sentiment. A regression line always has an error term because, in real life, independent variables are never perfect predictors of the dependent variables. Rather the line is an estimate based on the available data.



The distance between each point and the linear graph (shown as black arrows on the above graph) is our error term. So we can write our function as  $R^B = \beta_0 + \beta_1 E^x + \varepsilon$  where  $\beta_0$  and  $\beta_1$  are constants and  $\varepsilon$  is an (non constant) error term.