Regression and Classification

Regression: Regression is the process of predicting a Label(or Dependent Variable) based on the features(Independent Variables). It is used for modelling and finding the causal effect relationship between the variables and forecasting, where expected forecast is continuous in nature. For example, the relationship between the stock prices of the company and various factors like customer reputation and company annual performance etc. can be studied using regression.

Classification: It is the process of modelling and finding the causal effect relationship between the variables and forecasting, where expected forecast is class in nature. For example to predicting the expected output as dog or cat, man or woman, male or female etc., can be studied using classification.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Linear Regression

Linear Regression is one of the most fundamental algorithms in the Machine Learning world.

Building blocks of a Linear Regression Model are:

- Discreet/continuous independent variables
- A best-fit regression line
- Continuous dependent variable. i.e., A Linear Regression model predicts the dependent variable using a regression line based on the independent variables. The equation of the Linear Regression is:

$$Y=m*X+c+e$$

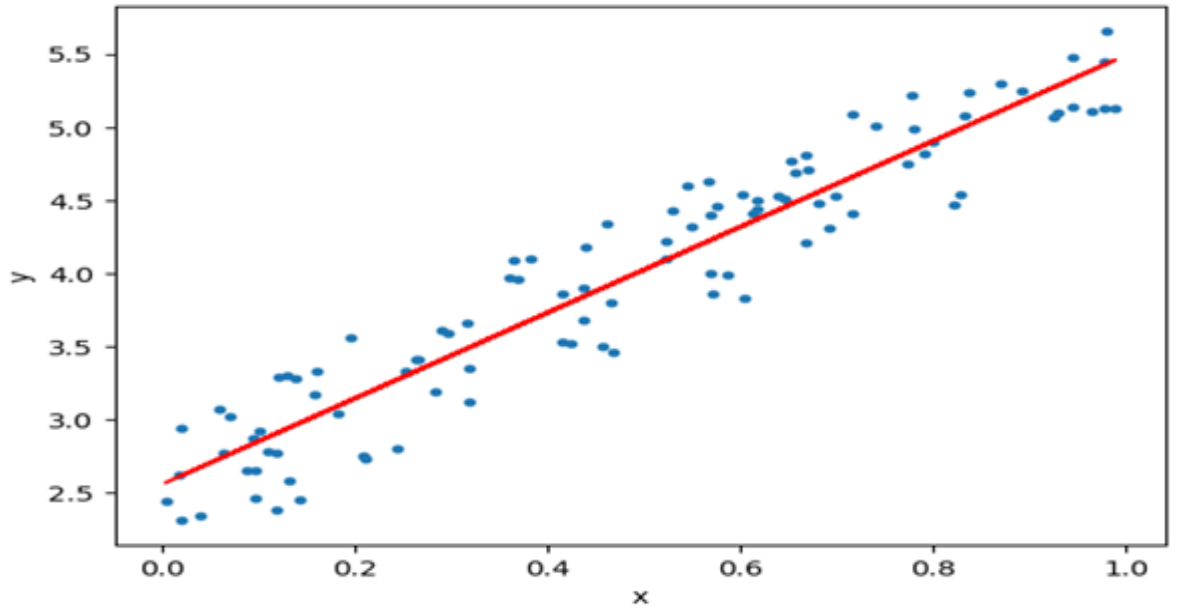
Where, c is the intercept, m is the slope of the line, and e is the error term. The equation above is used to predict the value of the target variable based on the given predictor variable(s).

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

When to use Linear Regression?

If the relation between independent and dependent variables are linear in nature then Linear Regression is used to solve the problem.

In the below picture blue dots are the distribution of ' dependent variable(y) ' w.r.t independent variable (x) and the data shows linear trend and Linear Regression can be used to solve this problem.

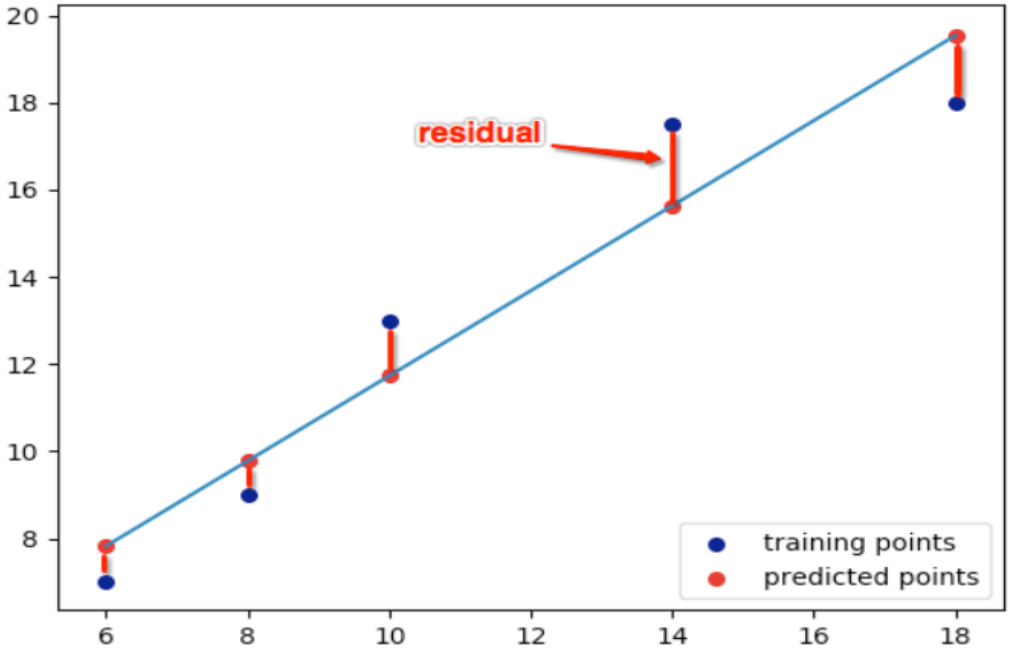


Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is Residual?

Residual is the distance between the actual output(Y) and the predicted output(Ŷ), as shown below picture.

Mathematically, Residual is: $r=y-(mx+b)$



Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Plots used to showcase the relationship amongst Independent Variables

- Scatterplots and Heatmap

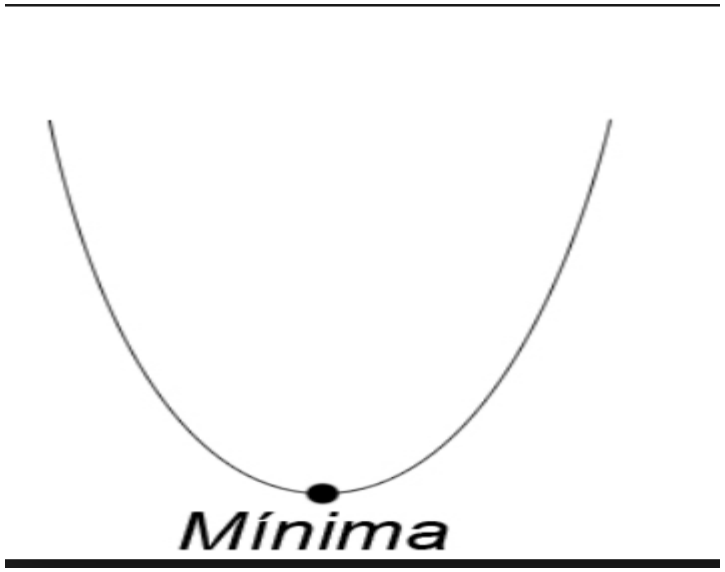
How is the best fit line chosen?

- The best fit line is obtained by minimizing the residual.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is gradient descent, and why is it used?

Gradient descent is a first-order iterative optimization algorithm for finding a local minimum of a differentiable function. To find a local minimum of a function using gradient descent, we take steps proportional to the negative of the gradient of the function at the current point. It is a technique to find the point where the error is as minimum as possible.
Pictorial representation of gradient descent can be like:



Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Mathematics Behind gradient descent:

Gradient Descent is use to find the best fit line i.e minimum residual.
Mathematically, Residual is: $r=y-(mx+b)$
Hence, the sum of the square of residuals is:

$$r_i = y_i - (mx_i + b) \quad \text{(Residual for one point)}$$

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} (y_i - (mx_i + b)) \quad \text{(Sum of residuals)}$$

$$R(x) = \sum_{i}^{n} r_i^2 = \sum_{}^{n} (y_i - (mx_i + b))^2 \quad \text{(Sum of squares of residuals)}$$

As we can see that the residual is both a function of m and b, so differentiating partially with respect to m and b will give us:

$$\frac{\partial R}{\partial m} = \sum_{i=0}^{n} 2x_i (b + mx_i - y_i)$$

$$\frac{\partial R}{\partial b} = \sum_{i=0}^{n} 2 (b + mx_i - y_i)$$

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

For getting the best fit line, residual should be minimum. The minima of a function occurs where the derivative=0. So, equating our corresponding derivatives to 0, we get:

$$\sum_{i=0}^{n} 2x_i\,(b + mx_i - y_i) = 0$$

$$\sum_{i=0}^{n} 2\,(b + mx_i - y_i) = 0$$

$$-$$

$$\sum_{i=0}^{n} 2x_i b + 2mx_i^2 - 2y_i x_i = 0$$

$$\sum_{i=0}^{n} 2b + 2mx_i - 2y_i = 0$$

$$-$$

$$\sum_{i=0}^{n} 2x_i b + \sum_{i=0}^{n} 2mx_i^2 - \sum_{i=0}^{n} 2y_i x_i = 0$$

$$\sum_{i=0}^{n} 2b + \sum_{i=0}^{n} 2mx_i - \sum_{i=0}^{n} 2y_i = 0$$ (Break up the summations)

$$-$$

$$\sum_{i=0}^{n} x_i b + \sum_{i=0}^{n} mx_i^2 - \sum_{i=0}^{n} y_i x_i = 0$$

$$\sum_{i=0}^{n} b + \sum_{i=0}^{n} mx_i - \sum_{i=0}^{n} y_i = 0$$ (diving both sides by 2)

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

In General, values for 'slope' and 'intercept' are calculated as follows:

$$\text{repeat until convergence } \{$$
$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)$$
$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) \cdot x^{(i)}$$
$$\}$$

where, θ_0 is 'intercept' , θ_1 is the slope, α is the learning rate, m is the total number of observations and the term after the $\sum$ sign is the loss.

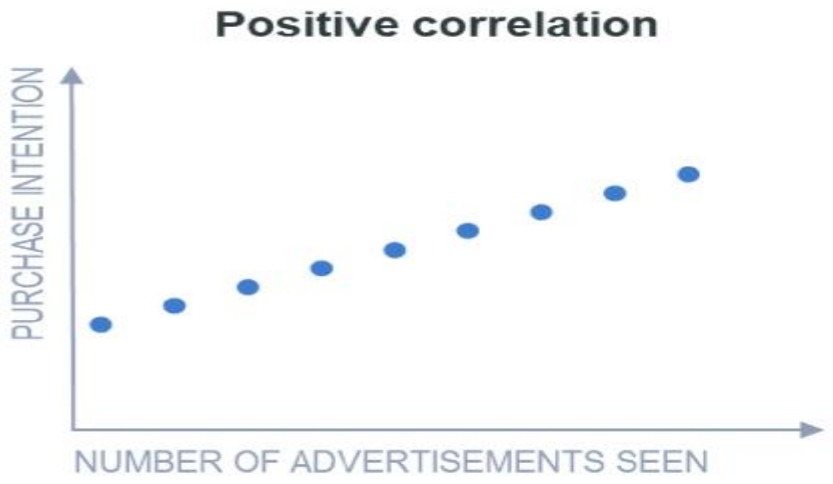
Google Tensor board recommends a Learning rate between 0.00001 and 10. Generally a smaller learning rate is recommended to avoid overshooting while creating a model.

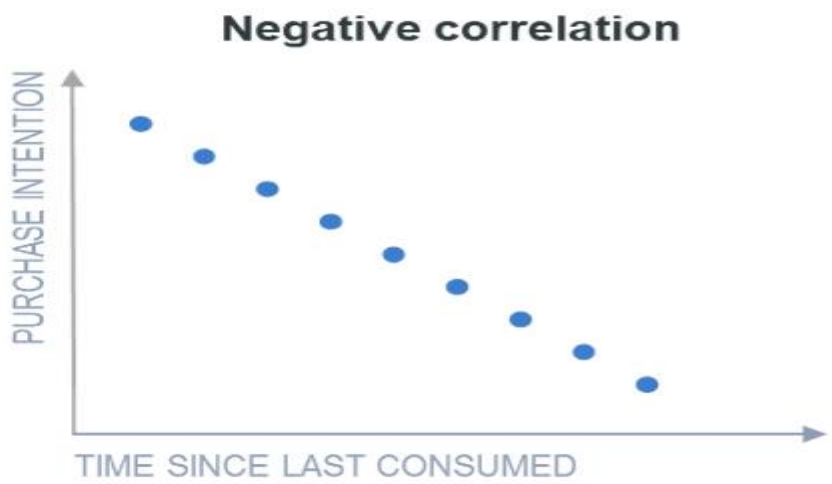
Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is correlation?

Correlation is a measure of the strength of a linear relationship between two quantitative variables.
Correlation is broadly classified as positive and negative Correlation.

Positive correlation is a relationship between two variables in which both variables move in the same direction. This is when one variable increases while the other increases and visa versa. Negative correlation is a relationship where one variable increases as the other decreases, and vice versa.





Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is Multicollinearity?

Multicollinearity is a problem which arises due the correlation between independent variables in a regression model. This correlation is a problem because independent variables should be independent. If the degree of correlation between variables is very high, it can cause problems in fitting the model and interpret the results.

How to detect Multicollinearity?

Multicollinearity can be detected by VIF(Variance Inflation factor). VIF is the factor of the variance in a model. It quantifies the severity of Multicollinearity. VIF of less than 5 is considered as no Multicollinearity, but this is just recommendation. It's not necessary that if VIF is less than 5 then no Multicollinearity.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What are the remedies for Multicollinearity?

Multicollinearity can treated with following methods:

- Remove some of the highly correlated independent variables.
- Linearly combine the independent variables, such as adding them together.
- Perform an analysis designed for highly correlated variables, such as principal components analysis or partial least squares regression.

How to interpret a Linear Regression model?

The equation for Linear Regression is as below: $y=\beta 0+\beta 1x$
$\beta 0$ - intercept $\beta 1$ - coefficient

For example: If $\beta 1$ = 0.047537 A "unit" increase in independent variable(x) is associated with a 0.047537 "unit" increase independent variables(y) Or,an additional $1,000 on independent variable(x) is translated to an increase in independent variables(y)by 47.53 Dollars. As an increase in x is associated with a decrease in y, $\beta 1$ would be negative.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is the R-Squared Statistics?

The R-squared statistic provides a measure of fit. It takes the form of a proportion, the proportion of variance explained.
It always takes on a value between 0 and 1.
Where RSS: is the Residual Sum of squares and is given as :

In simple words, it represents how much of our data is being explained by our model. For example, $R2$ statistic = 0.75, it says that our model fits 75 % of the total data set. Similarly, if it is 0, it means none of the data points is being explained and a value of 1 represents 100% data explanation. Mathematically $R2$ statistic is calculated as :

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Where RSS: is the Residual Sum of squares, the difference between expected and predicted result . It can be represented as:

$$\text{RSS} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

TSS: is the Total sum of squares, the difference between expected output and it's mean. It can be represented as: :

$$\text{TSS} = \sum (y_i - \bar{y})^2$$

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is an adjusted R-Squared Statistics?

With increase in number of independent variables, the R2 increases as well. But that doesn't mean that the new independent variables have any correlation with the output variable.
In other words, even with the addition of new features in our model, it is not necessary that our model will yield better results but R2 value will increase.
To rectify this problem, we use Adjusted R2 value which penalizes excessive use of such features which do not correlate with the output data. Mathematically, it is calculated as:

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

where

R^2 = sample R-square

p = Number of predictors

N = Total sample size.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is the difference between fit, fit_transform and predict methods?

1. fit method is used for training of model using data. fit computes the mean and std to be used for later scaling (just a computation), nothing is given as output.

2. fit_transform uses a previously computed mean and std to auto scale the data (subtract mean from all values and then divide it by std). fit_transform is used to change data to polynomial.

3. predict method is use to test the prediction capability of trained model. This method is performed on a dataset to predict the response variable based study the relationship between a response and predictor variable

How do you plot the least squared line?

Least squared line is plotted between the test data and the predicted output by model. For example if x is the data and the predicted value by model is p then the least squared line is plotted between x and p.

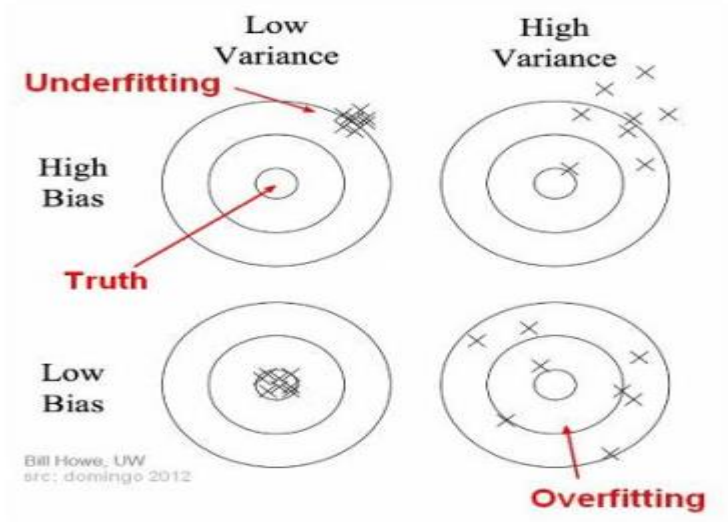
Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What are Bias and Variance?

Bias: Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance: Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

Bias and variance using bulls-eye diagram:



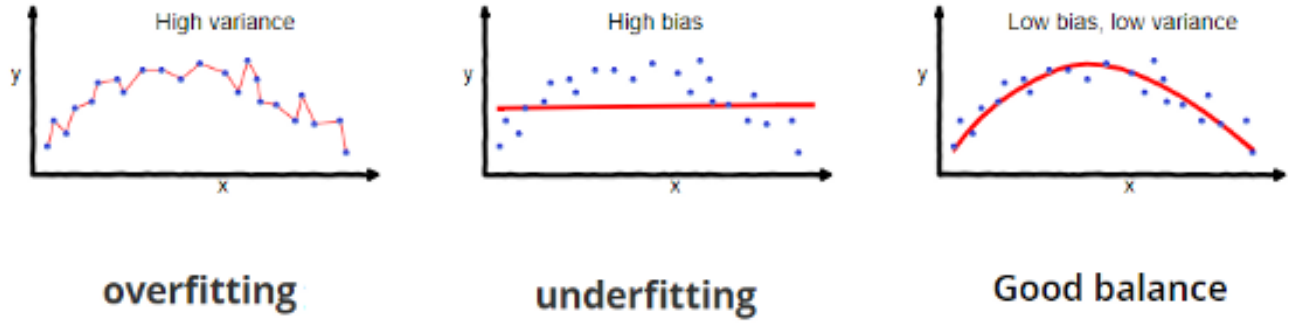
Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What are Bias and Variance?

In the bull eye diagram, center of the target is a model that perfectly predicts correct values. As we move away from the bulls-eye our predictions become get worse and worse. We can repeat our process of model building to get separate hits on the target.

In supervised learning, **underfitting** happens when a model unable to capture the underlying pattern of the data. These models usually have high bias and low variance. It happens when we have very less amount of data to build an accurate model or when we try to build a linear model with a nonlinear data. Also, these kind of models are very simple to capture the complex patterns in data like Linear and logistic regression.

In supervised learning, **overfitting** happens when our model captures the noise along with the underlying pattern in data. It happens when we train our model a lot over noisy dataset. These models have low bias and high variance. These models are very complex like Decision trees which are prone to overfitting.



Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Bias Variance Trade-off

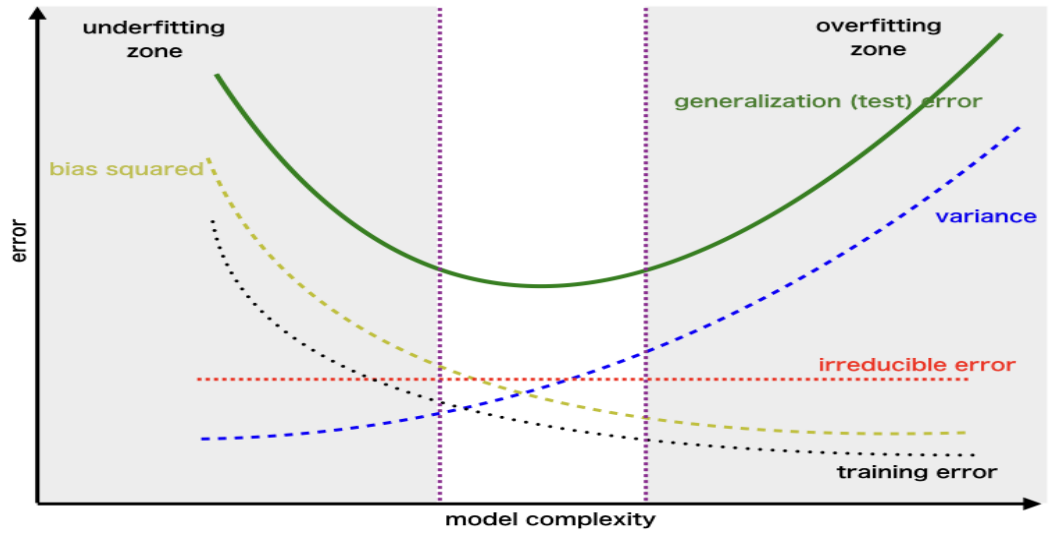
If the model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.

A tradeoff between bias and variance is necessary so that algorithm can't be more complex and less complex at the same time.

To build a good model, we need to find a good balance between bias and variance such that it minimizes the total error.

Total Error = Bias^2 + Variance + Irreducible Error

The error to complexity graph to show trade-off is given as:–



Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is the null and alternate hypothesis?

Null Hypothesis: Null Hypothesis is the hypothesis to be tested for possible rejection under the assumption that it is true. The concept of the null is similar to innocent until proven guilty We assume innocence until we have enough evidence to prove that a suspect is guilty. It is denoted by H0.

Alternate Hypothesis : Alternate Hypothesis is the alternative hypothesis complements the Null hypothesis. It is opposite of the null hypothesis such that both Alternate and null hypothesis together cover all the possible values of the population parameter. It is denoted by H1.

For example: A soap company claims that it's product kills on an average 99% of the germs. To test the claim of this company we will formulate the null and alternate hypothesis. Null Hypothesis(H0): Average =99% Alternate Hypothesis(H1): Average is not equal to 99%.

What is multiple linear regression?

The multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables. The independent variables can be continuous or categorical.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is the OLS(Ordinary Least Squared) method?

The OLS is an strategy to obtain, a ‘straight line', from model, which is as close as possible to the data points. OLS minimize the squared errors to save the model to penalize by compensating the positive and negative errors.

The equation for the Linear Regression model is Y = mX+b, where m is the slope and b is the intercept.

Mathematically, Residual is: $r=y-(mx+b)$ and the Ordinary Least Squared error is derived as:

$$r_i = y_i - (mx_i + b) \quad \text{(Residual for one point)}$$

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} (y_i - (mx_i + b)) \quad \text{(Sum of residuals)}$$

$$R(x) = \sum_{i}^{n} r_i^2 = \sum (y_i - (mx_i + b))^2 \quad \text{(Sum of squares of residuals)}$$

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is the p-value? How does it help in feature selection?

While performing a hypothesis test, a p-value helps in determining the significance of results. Hypothesis tests are used to test the validity of a claim that is made about a population. This claim that's on trial is called the null hypothesis.

The alternative hypothesis is the one which tells the null hypothesis is concluded to be false. All hypothesis tests ultimately use a p-value to weigh the strength of the evidence. The p-value is a number between 0 and 1 and interpreted in the following way:

A small p-value (typically ≤ 0.05) indicates strong evidence against the null hypothesis, so you reject the null hypothesis.

A large p-value (> 0.05) indicates weak evidence against the null hypothesis, so you fail to reject the null hypothesis.

p-values very close to the cutoff (0.05) are considered to be marginal (could go either way). Always report the p-value so your readers can draw their own conclusions.

Using 0.05 as the cutoff is just a convention.

How does it help in feature selection?

The p-value represents the probability of the coefficient actually being zero.

If the 95% confidence interval includes zero, the p-value for that coefficient will be greater than 0.05. If the 95% confidence interval does not include zero, the p-value will be less than 0.05.

p-value of less than 0.05 shows that there is some relationship between the feature selected and response(dependent variable).

The p-value for the intercept is generally ignored.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

How to handle categorical values in the data?

Categorical values in the data can be handled in following ways:

1.Label Encoding: In this method each label is assigned a unique integer.

2.One-Hot Encoding: It creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. One-Hot Encoding is the process of creating dummy variables.

What is regularization?

When we use regression models to train some data, there is a good chance that the model will over fit the given training data set. Regularization helps sort this overfitting problem by restricting the degrees of freedom of a given equation i.e. simply reducing the number of degrees of a polynomial function by reducing their corresponding weights.
In a linear equation, we do not want huge weights/coefficients as a small change in weight can make a large difference for the dependent variable (Y). So, regularization constraints the weights of such features to avoid overfitting.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Why to use Regularization?

Regularization helps to reduce the variance of the model, without a substantial increase in the bias. If there is variance in the model that means that the model won't fit well for dataset different that training data.

The tuning parameter λ controls this bias and variance tradeoff. When the value of λ is increased up to a certain limit, it reduces the variance without losing any important properties in the data. But after a certain limit, the model will start losing some important properties which will increase the bias in the data.

Thus, the selection of good value of λ is the key. The value of λ is selected using cross-validation methods. A set of λ is selected and cross-validation error is calculated for each value of λ and that value of λ is selected for which the cross-validation error is minimum.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Ridge Regression

Ridge Regression (L2 Form) Ridge regression penalizes the model based on the sum of squares of magnitude of the coefficients.
The regularization term is given by regularization,

$$\lambda * \sum |\beta_j^2|$$

Where, λ is the shrinkage factor.
and hence the formu

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

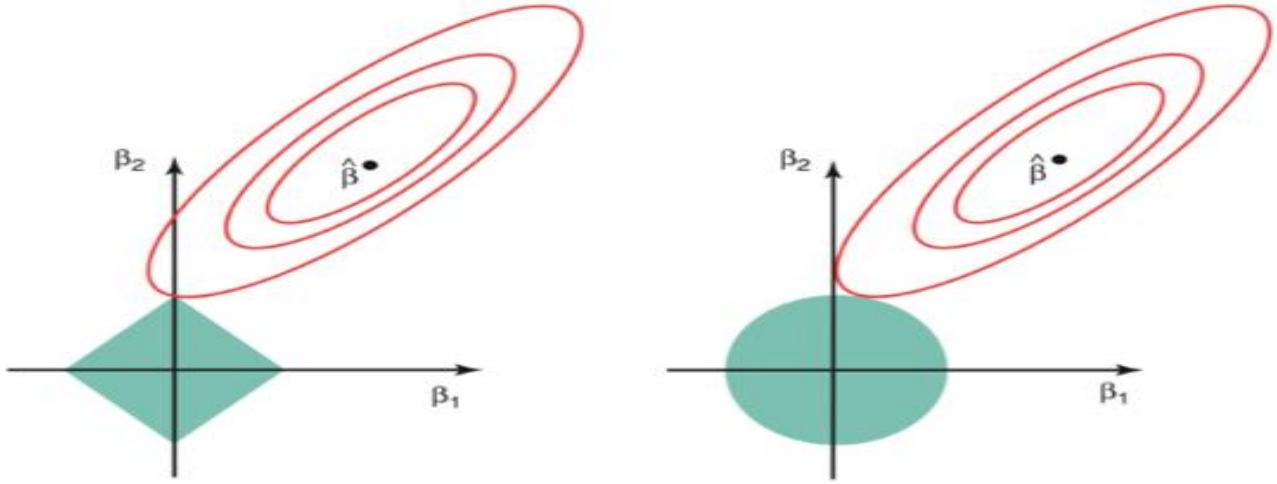
This value of lambda can be anything and should be calculated by cross validation as to what suits the model.
Let's consider β1 and β2 be coefficients of a linear regression and λ = 1:

For Lasso, $\beta_1 + \beta_2 <= s$

For Ridge, $\beta_1^2 + \beta_2^2 <= s$

Where s is the maximum value the equations can achieve . If we plot both the above equations, we get the following graph:

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar



Lasso Regression

The red ellipse represents the cost function of the model, whereas the square (left side) represents the Lasso regression and the circle (right side) represents the Ridge regression.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Lasso Regression

LASSO(Least Absolute Shrinkage and Selection Operator) Regression (L1 Form)
LASSO regression penalizes the model based on the sum of magnitude of the coefficients. The regularization term is given by regularization,

$$\lambda * \sum |\beta_j|$$

Where, λ is the shrinkage factor.
and hence the formula for loss after regularization is:

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

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Elastic Net

According to the Hands-on Machine Learning book, elastic Net is a middle ground between Ridge Regression and Lasso Regression. The regularization term is a simple mix of both Ridge and Lasso's regularization terms, and you can control the mix ratio α.

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i-1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j-1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j-1}^{m}|\hat{\beta}_j|\right),$$

Why do we do a train test split?

The train test split of dataset is done to generalize the model behavior. If the model is fit on the training dataset, then it implicitly minimize error or find correct responses. The fitted model provides a good prediction on the training dataset. Then to test dataset is use to check the response of model with unknown data.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

What is polynomial regression?

For understanding Polynomial Regression, let's first understand a polynomial. Merriam-webster defines a polynomial as: "*A mathematical expression of one or more algebraic terms each of which consists of a constant multiplied by one or more variables raised to a non-negative integral power (such as a + bx + cx^2)".* Simply said, poly means many. So, a polynomial is an aggregation of many monomials(or Variables).

A simple polynomial equation can be written as:

$y=a+bx+cx2+...+nxn+...$y=a+bx+cx2+...+nxn+...

So, Polynomial Regression can be defined as a mechanism to predict a *dependent variable* based on the polynomial relationship with the *independent variable*.
In the equation,

$y=a+bx+cx2+...+nxn+...$y=a+bx+cx2+...+nxn+...

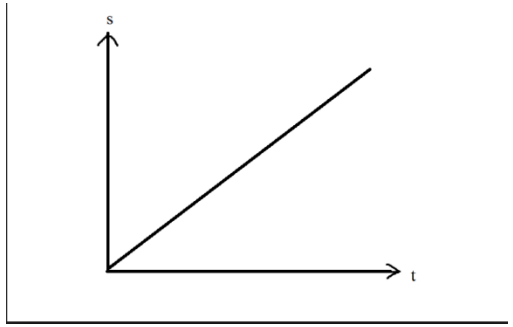
the maximum power of 'x' is called the degree of the polynomial equation. For example, if the degree is 1, the equation becomes $y=a+bx$y=a+bx
which is a simple linear equation.

if the degree is 2, the equation becomes
$y=a+bx+cx2$y=a+bx+cx2
which is a quadratic equation and so on.

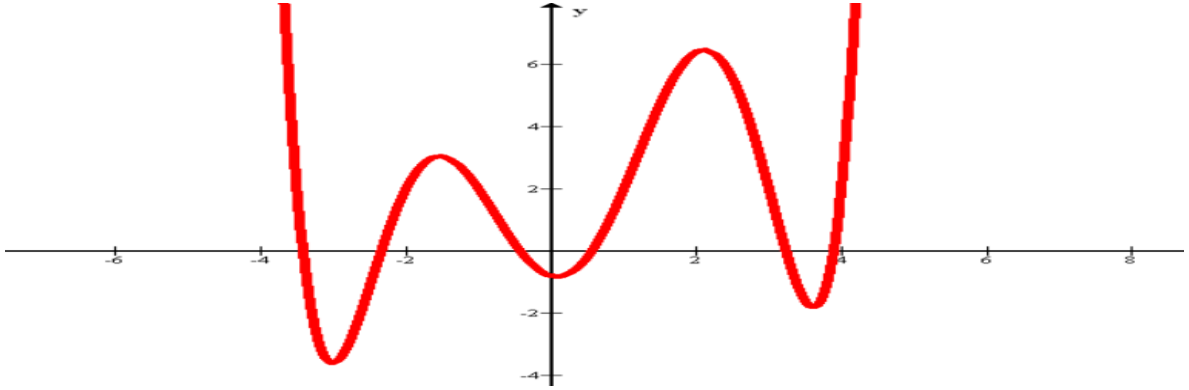
Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

When to use Polynomial Regression?

Many times we may face a requirement where we have to do a regression, but when we plot a graph between a dependent and independent variables, the graph doesn't turn out to be a linear one. A linear graph typically looks like:



But what if the relationship looks like:



Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

It means that the relationship between X and Y can't be described Linearly. Then comes the time to use the Polynomial Regression.
We can generalize the matrix obtained above (for Linear Regression) for an equation of n coefficients(in y=mx+b, m and b are the coefficients) as follows:

$$\begin{bmatrix} n & \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \cdots & \sum_{i=0}^{n} x_i^m \\ \sum_{i=0}^{n} x_i & \sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i^3 & \cdots & \sum_{i=0}^{n} x_i^{(m+1)} \\ \sum_{i=0}^{n} x_i^2 & \sum_{i=0}^{n} x_i^3 & \sum_{i=0}^{n} x_i^4 & \cdots & \sum_{i=0}^{n} x_i^{(m+2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=0}^{n} x_i^m & \sum_{i=0}^{n} x_i^{(m+1)} & \sum_{i=0}^{n} x_i^{(m+2)} & \cdots & \sum_{i=0}^{n} x_i^{2m} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} = \begin{bmatrix} \sum_{i=0}^{n} y_i \\ \sum_{i=0}^{n} x_i y_i \\ \sum_{i=0}^{n} x_i^2 y_i \\ \vdots \\ \sum_{i=0}^{n} x_i^m y_i \end{bmatrix}$$

Where m is the _degree_(maximum power of x) of the polynomial and n is the number of observation points. The above matrix results in the general formula for Polynomial Regression. Earlier, we were able to visualize the calculation of minima because the graph was in three dimensions. But as there are n number of coefficients, it's not possible to create an (n+1) dimension graph here.

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Label Encoding and One Hot Encoding

Both Label Encoding and One Hot Encoding are way of handling categorical data.

Label Encoding: With label encoding, numbers are assigned to different categorical data. For example, if we are dealing with two categories Dog and Cat. Label encoding means assigning number like 1 to Dog and 2 to Cat, so that computer will be able to under.

Problem with Label Encoding:
The problem is that with label encoding, the categories now have natural ordered relationships. The computer does this because it's programmed to treat higher numbers as high in order, it will give higher weights to higher number.

One Hot Encoding:
It is binary style of converting categorical data to numerical data.

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar

Label Encoding and One Hot Encoding

One Hot Encoding:
It is binary style of converting categorical data to numerical data. For example:

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50

Limitation of Scikit learn Library:
Sklearn's one hot encoder (OneHotEncoder) doesn't know how to convert categories to numbers, it only knows how to convert numbers to binary. So LabelEncoder need to be used first to assign number to categorical data then OneHotEncoder should be used to change it into binary format.
Both LabelEncoder and OneHotEncoder methods need to required to use One Hot Encoding

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

Author: Harshbardhan Kumar
Linkedin : www.linkedin.com/in/harshbardhan-kumar