GATE in Data Science & Artificial Intelligence

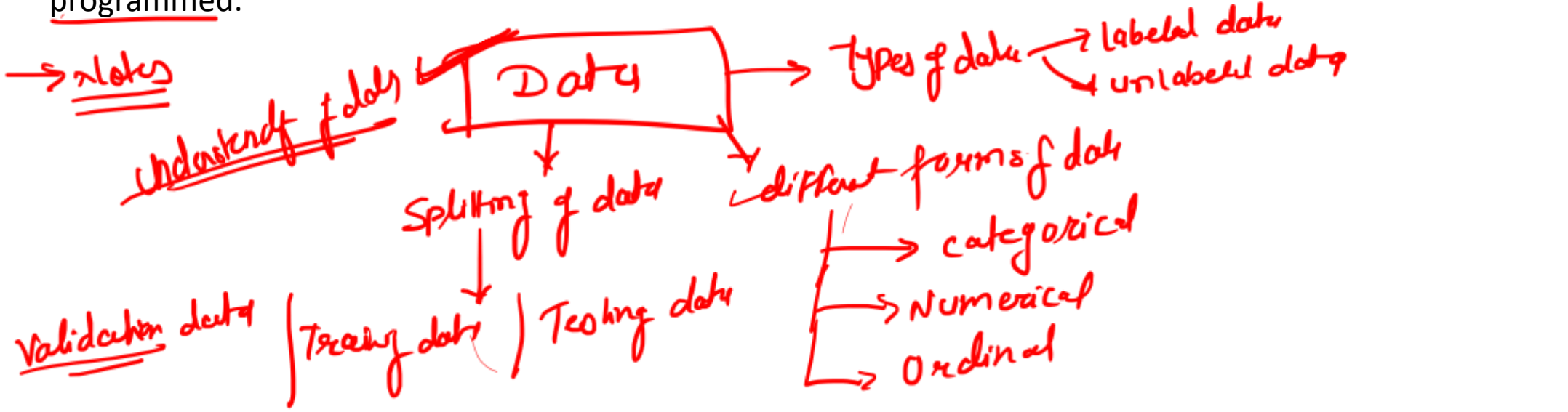
Machine Learning

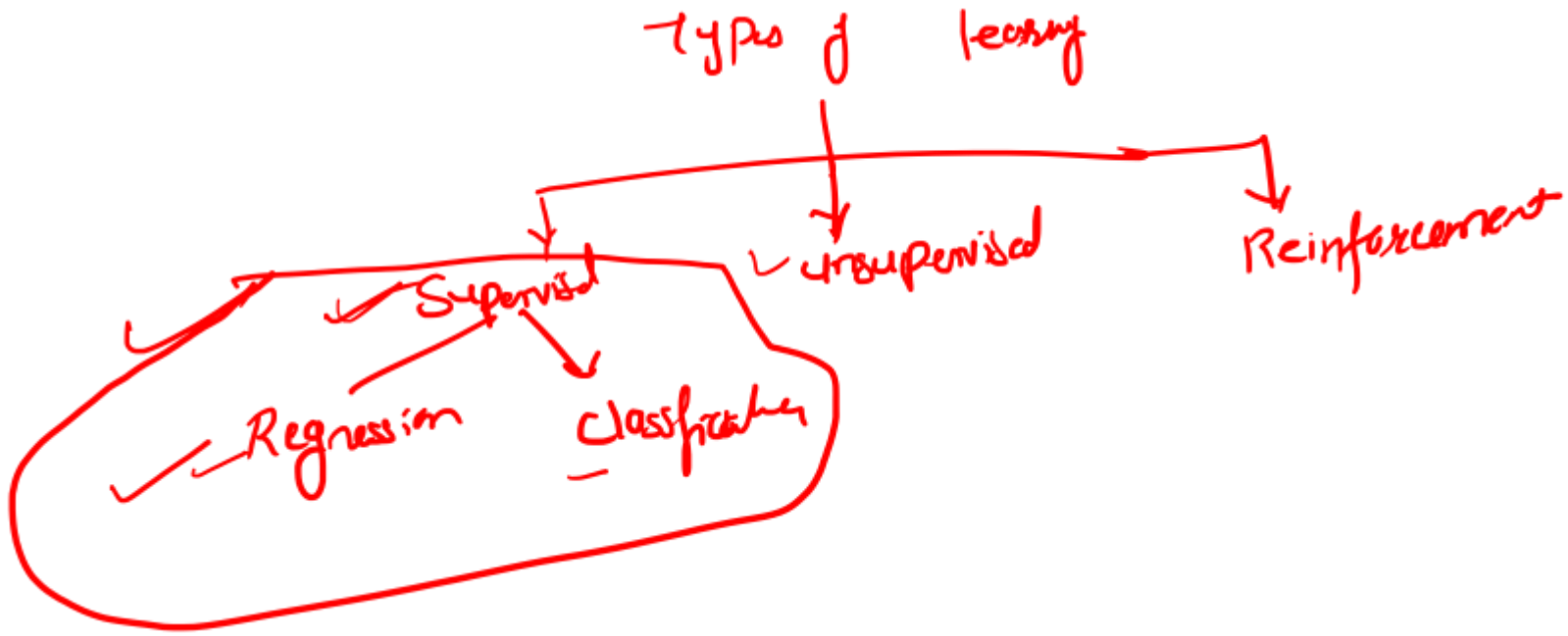
Intro to ML

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree

Machine Learning: (i) Supervised Learning: regression and classification problems, simple linear regression, multiple linear regression, ridge regression, logistic regression, k-nearest neighbour, naive Bayes classifier, linear discriminant analysis, support vector machine, decision trees, bias-variance trade-off, cross-validation methods such as leave-one-out (LOO) cross-validation, k-folds cross-validation, multi-layer perceptron, feed-forward neural network; (ii) Unsupervised Learning: clustering algorithms, k-means/k-medoid, hierarchical clustering, top-down, bottom-up: single-linkage, multiple-linkage, dimensionality reduction, principal component analysis.

Machine Learning is the field of study that gives computers the capability to learn without being explicitly programmed.



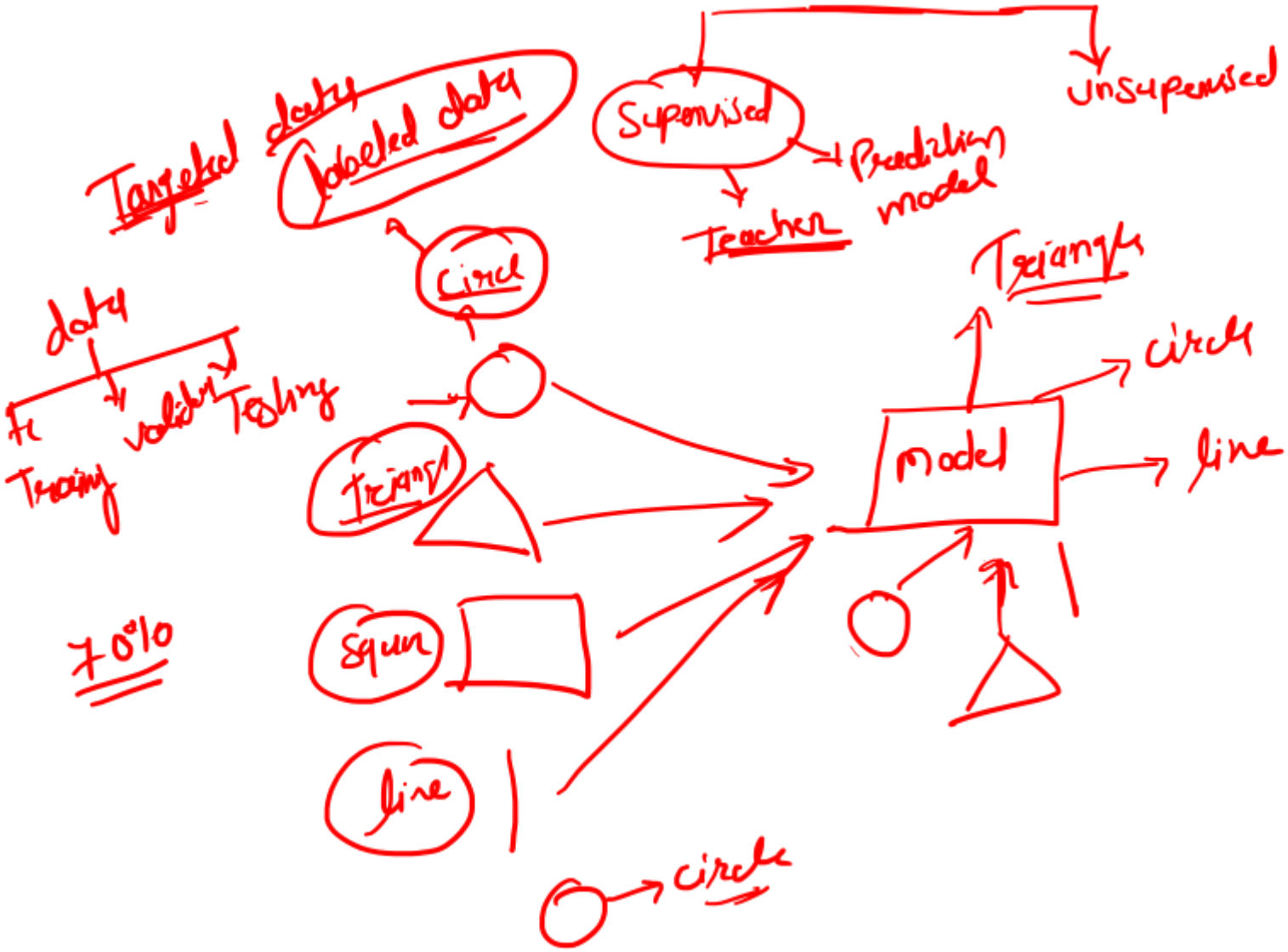

Types of learning
Supervised
Unsupervised
Reinforcement
Regression
Classification

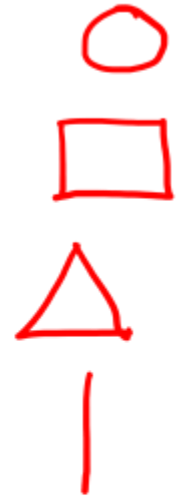
GATE in Data Science & Artificial Intelligence

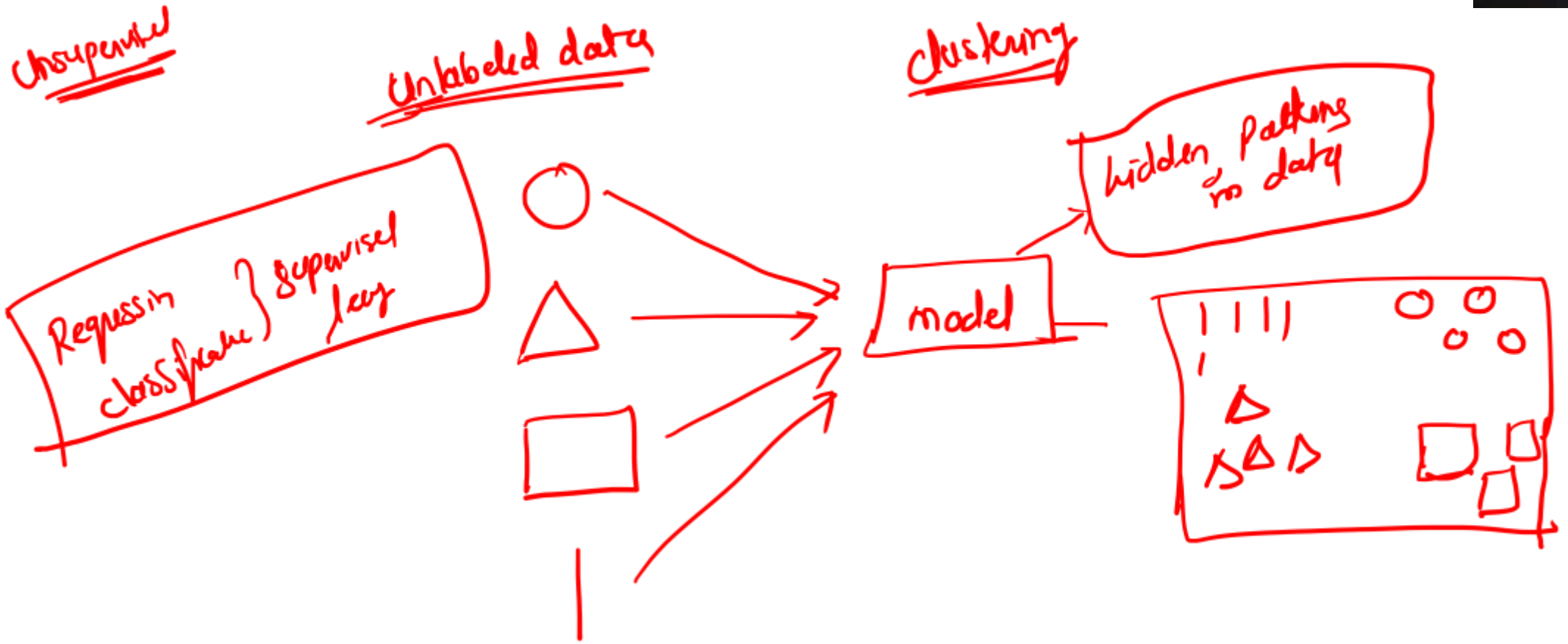
Machine Learning

Supervised Vs Unsupervised Learning

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree


Targeted data
labeled data
Supervised
Unsupervised
Prediction model
Teacher
data
Testing
Training
circle
Triangle
square
line
model
70%
circle




Unlabeled data
clustering
Regression
classification
Supervised learning
model
hidden patterns in data

GATE in Data Science & Artificial Intelligence

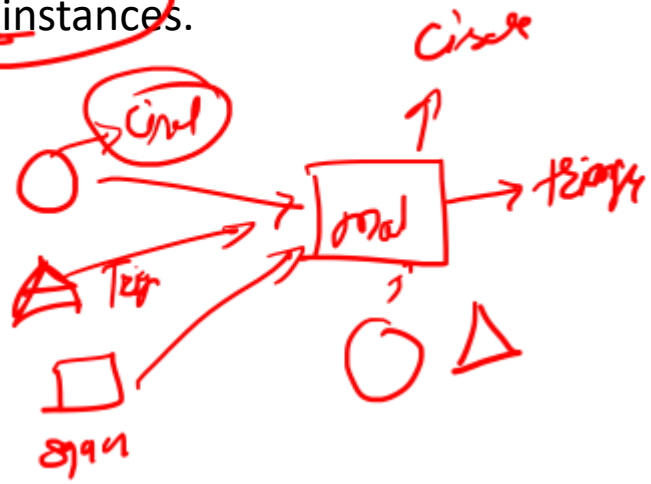
Machine Learning

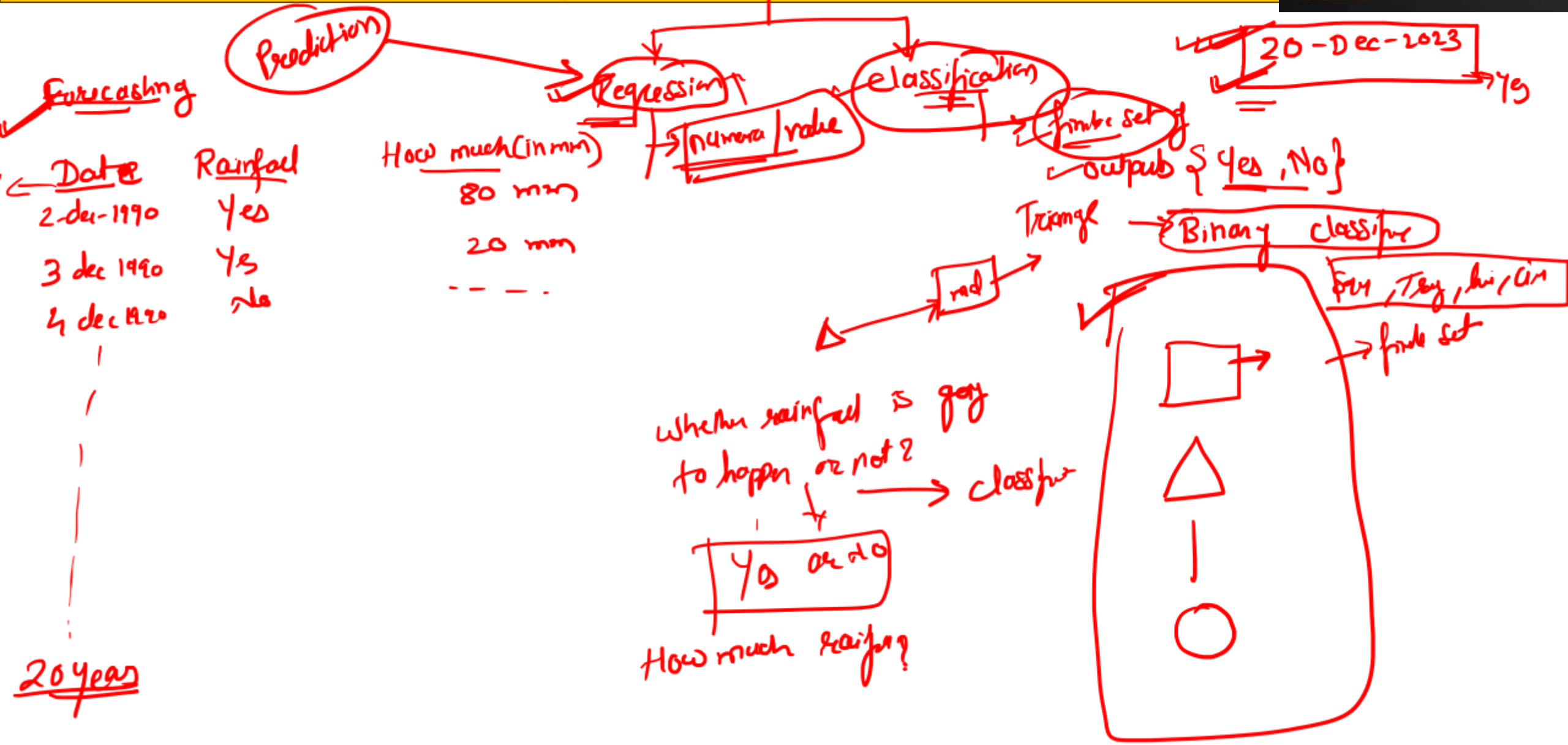
Supervised Learning

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree

Supervised Learning

- Supervised learning is the machine learning task of learning a function that maps an input to an output based on example input-output pairs.
- In supervised learning, each example in the training set is a pair consisting of an input object (typically a vector) and an output value.
- A supervised learning algorithm analyzes the training data and produces a function, which can be used for mapping new examples.
- In the optimal case, the function will correctly determine the class labels for unseen instances.
- Both classification and regression problems are supervised learning problems





Classification & Regression

- Predicting the house price.
- Predicting whether it will rain or not on a given day.
- Predicting the maximum temperature on a given day.
- Predicting the sales of the ice-creams.
- Predicting whether a patient is diagnosed with cancer or not.
- Predicting whether a team will win a tournament or not.
- Predicting the price of a second-hand car.
- Classify web text into one of the following categories: Sports, Entertainment, or Technology.

Performance metric for least square regression Performance metric for least square regression Performance metric for least square regression

GATE in Data Science & Artificial Intelligence

Regression vs classifn

Machine Learning

Regression

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree

- In machine learning, a regression problem is the problem of predicting the value of a numeric variable based on observed values of the variable.
- The value of the output variable may be a number, such as an integer or a floating point value. These are often quantities, such as amounts and sizes. The input variables may be discrete or real-valued.
- Regression algorithms are used if there is a relationship between the input variable and the output variable.
- It is used for the prediction of continuous variables, such as Weather forecasting, Market Trends, etc.

$y = f(x, \theta)$

General Approach

Let x denote the set of input variables and y the output variable. In machine learning, the general approach to regression is to assume a model, that is, some mathematical relation between x and y, involving some parameters say, θ, in the following form:

$$y = f(x, \theta)$$

The function $f(x, \theta)$ is called the regression function. The machine learning algorithm optimizes the parameters in the set θ such that the approximation error is minimized; that is, the estimates of the values of the dependent variable y are as close as possible to the correct values given in the training set.

$\hat{y}$

$$\text{Residual error} = y - \hat{y}$$

y

Various types of regression techniques

These techniques mostly differ in three aspects, namely, the number and type of independent variables, the type of dependent variables and the shape of regression line. Some of these are listed below.

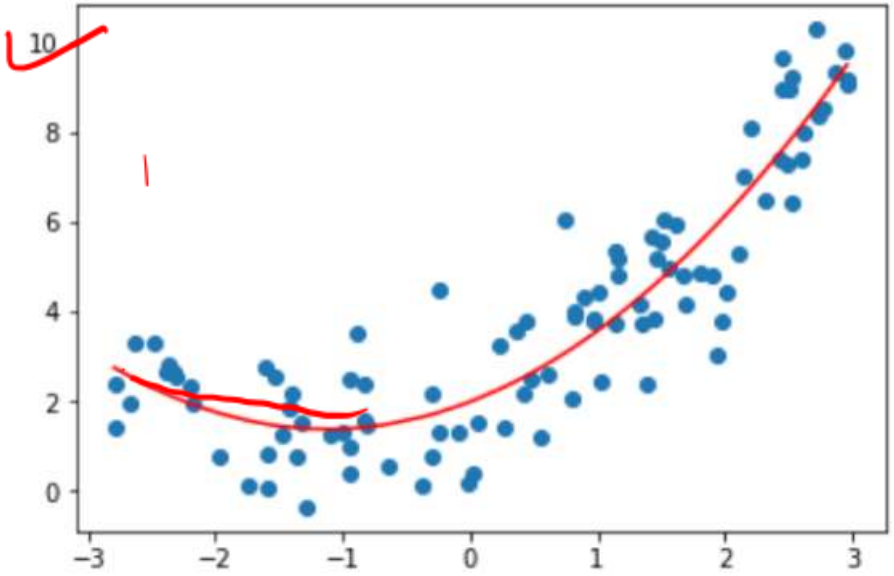
1. Simple linear regression: There is only one continuous independent variable x and the assumed relation between the independent variable and the dependent variable y is

$$y = a + bx$$

2. Multivariate linear regression: There are more than one independent variable, say $x1, ..., xn$, and the assumed relation between the independent variables and the dependent variable is $y = a_0 + a_1x_1 + ... + a_nx_n$

3. Polynomial regression: There is only one continuous independent variable x and the assumed model is $y = a_0 + a_1 x + ... + a_n x^n$
It is a variant of the multiple linear regression model, except that the best fit line is curved rather than straight.

$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3$



4. Ridge regression: Ridge regression is one of the types of linear regression in which a small amount of bias is introduced so that we can get better long-term predictions. Ridge regression is a regularization technique, which is used to reduce the complexity of the model. It is also called as L2 regularization.

5. Logistic regression: The dependent variable is binary, that is, a variable which takes only the values 0 and 1. The assumed model involves certain probability distributions.

GATE in Data Science & Artificial Intelligence

Machine Learning

Simple Linear Regression

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree

$y = f(x, \theta)$ → General for Regression

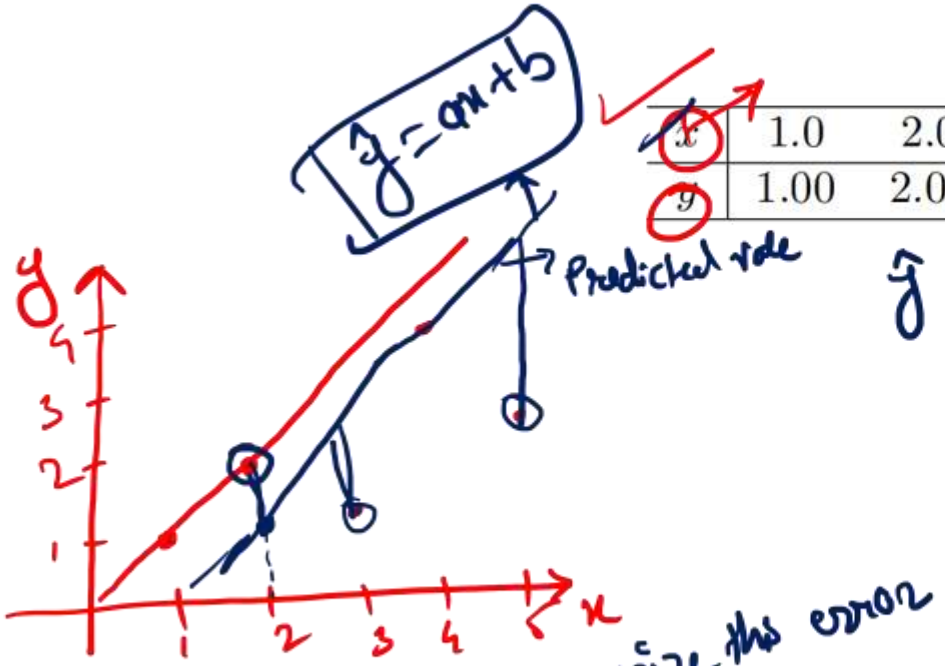
$y = ax + b$ → eqn of st line

a, b → regression coeffs

⇒ Best fit line

$\hat{y} = ax + b$

⇒ least square method

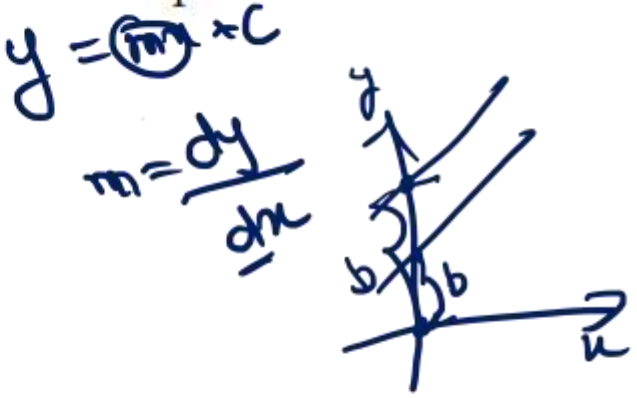


x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

6

Residual error $= |y - \hat{y}|$ ⇒ minimize this error

Simple linear regression is a basic machine learning technique used for modeling the relationship between a single independent variable (often denoted as "x") and a dependent variable (often denoted as "y"). It assumes a linear relationship between the variables and aims to find the best-fitting line (typically represented by the equation $y = mx + b$) that minimizes the sum of squared differences between the observed data points and the values predicted by the model.



Equation: The linear regression model is represented by the equation

$$y = ax + b$$

where:

y is the dependent variable (the one you want to predict).

x is the independent variable (the one used for prediction).

a is the slope (also called the regression coefficient), representing how much y changes for each unit change in x.

b is the y-intercept, representing the value of y when x is 0.

Formulas to find a and b

The means of x and y are given by

$$\bar{x} = \frac{1}{n}\sum x_i$$

$$\bar{y} = \frac{1}{n}\sum y_i$$

$$b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$$

$$a = \bar{y} - b\bar{x}$$

and also that the variance of x is given by

$$\text{Var}(x) = \frac{1}{n-1}\sum (x_i - \bar{x}_i)^2.$$

The covariance of x and y, denoted by Cov(x, y) is defined as

$$\text{Cov}(x, y) = \frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})$$

Obtain a linear regression for the data in below table assuming that y is the independent variable.

x	1.0	2.0	3.0	4.0	5.0
y	1.00	2.00	1.30	3.75	2.25

Stats

$x = ?$

$y = ax + b$

$y = bx + a$

$a = ?$, $b = ?$

$n = 5$

$\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{5}(1+2+3+4+5) = 3.0$

$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{5}(1+2+1.3+3.75+2.25) = 2.06$

$cov(x,y) = \frac{1}{n-1}\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right]$

$= \frac{1}{4}\left[(1-3)\cdot(1-2.06) + (2-3)\cdot(2-2.06) + \cdots + (5-3)\cdot(2.25-2.06)\right]$

$= 1.6625$

$Var(x) = \frac{1}{n-1}\sum (x_i - \bar{x})^2 = \frac{1}{4}\left[(1-3)^2 + (2-3)^2 + \cdots + (5-3.0)^2\right]$

$= 2.5$

$b = \frac{1.6625}{2.5} = 0.425$

$a = \bar{y} - b\bar{x} = 2.06 - 0.425 \times 3$

$a = 0.785$

$y = 0.425x + 0.785$

GATE in Data Science & Artificial Intelligence

Machine Learning

Multiple Linear Regression

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree

Multiple Linear Regression is a machine learning technique used to model the relationship between a dependent variable (target) and multiple independent variables (features) by fitting a linear equation to the data.

The model can be expressed as: $y = \beta_0 + \beta_1 x_1 + ... + \beta_N x_N$

Where:

y is the dependent variable (the one you want to predict).

$x_1, x_2, ..., x_n$ are the independent variables.

β_0 is the y-intercept.

$\beta_1, \beta_2, ...\beta_n$ are the coefficients associated with each independent variable.

Let there also be n observed values of these variables:

Variables (features)	Values (examples)			
	Example 1	Example 2	⋯	Example n
x_1	x_{11}	x_{12}	⋯	x_{1n}
x_2	x_{21}	x_{22}	⋯	x_{2n}
⋯				
x_N	x_{N1}	x_{N2}	⋯	x_{Nn}
y (outcomes)	y_1	y_2	⋯	y_n

As in simple linear regression, here also we use the ordinary least squares method to obtain the optimal estimates of $\beta_0, \beta_1, ...\beta_n$ The method yields the following procedure for the computation of these optimal estimates. Let

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{N1} \\ 1 & x_{12} & x_{22} & \cdots & x_{N2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{Nn} \end{bmatrix}, \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad B = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}$$

Then it can be shown that the regression coefficients are given by

$$B = (X^T X)^{-1} X^T Y$$

$\Rightarrow y = ax + b$ →

y_i = actual value at $x = i$

$\hat{y}_i$ = Predicted value at $x = i$

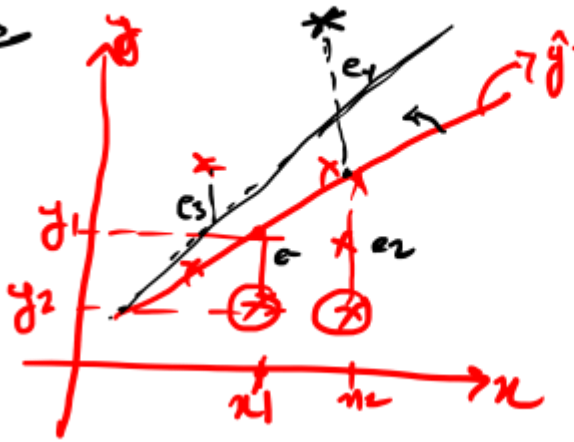
For x_1 →

error = $|y_1 - y_2|$

= |Actual value − Predicted value|

⇒ Rainfall (mm)

Outliers



$$\text{error} = \sum_{i=1} |y_i - \hat{y}_i|$$

(1) Mean Absolute error (MAE)

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

→ outlier effect will be less

→ Unit (mm)

(2) Mean square error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

→ outlier effect will be more

→ Unit will be different

mm^2

(3) Root mean square error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

→ outlier effect will be more

Unit (mm)

GATE in Data Science & Artificial Intelligence

R^2 & R^2_{adjust}

Machine Learning

MAE
MSE
RMSE

Linear Regression: Performance Metrics

By:
Piyush Wairale
MTech (IIT Madras)
Instructor at IIT Madras BS in Data Science Degree

$\Rightarrow R^2 \Rightarrow$ coeff of determination

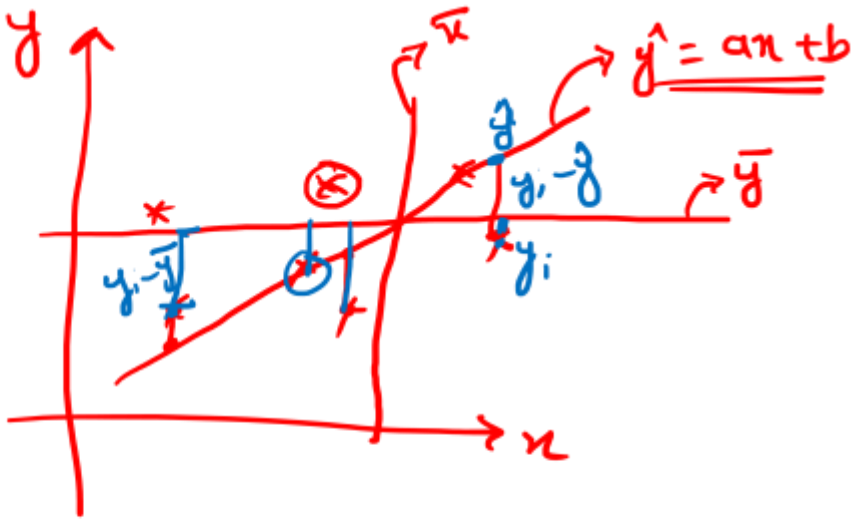
$y \Rightarrow$ actual value

$\hat{y} \Rightarrow$ Predicted value

$\bar{y}$ = mean of y.

$\hat{y} = ax + b$

$\sum (y_i - \hat{y}_i)^2$



$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}.$$

- The *coefficient of determination*, or R^2, is a measure that provides information about the goodness of fit of a model.
- In the context of regression it is a statistical measure of how well the regression line approximates the actual data.
- It is therefore important when a statistical model is used either to predict future outcomes or in the testing of hypotheses.

$R^2 \in [0, 1]$

ideal, $R^2 = 1$

$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

The *sum squared regression* is the sum of the residuals squared, and the *total sum of squares* is the sum of the distance the data is away from the mean all squared. As it is a percentage it will take values between 0 and 1.

Interpretation of the R^2 value

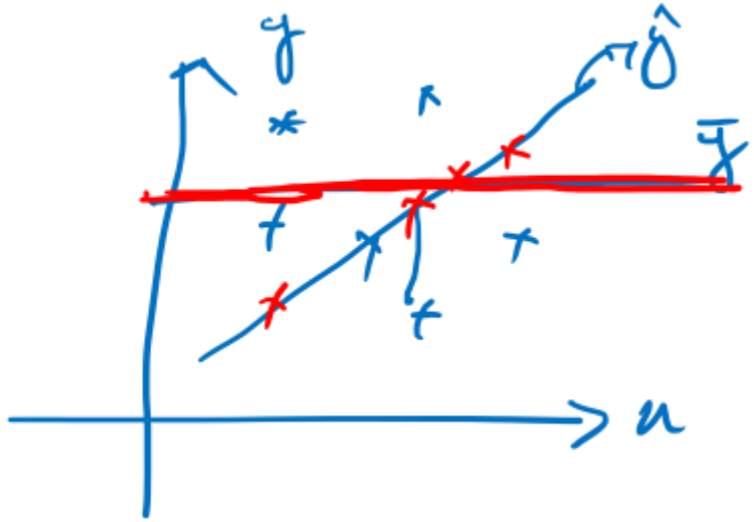
$$R^2 = 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}},$$

$$= 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}.$$

$\frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \to 1 \to 0$

$y_i = \hat{y}_i$ → ideal fit

when $R^2 = 1$, when $(y_i = \hat{y}_i)$



$(y_i - \hat{y}_i)^2 = (y_i - \bar{y})^2$

$\Rightarrow y_i + \hat{y}_i = y_i + \bar{y}$

$\Rightarrow \hat{y}_i = \bar{y}$

when $R^2 = 0$

$\Rightarrow$ higher the value of R^2, better will be my fit

The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable. In doing so, we can determine whether adding new variables to the model actually increases the model fit.

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

- **n** represents the number of data points in our dataset
- **k** represents the number of independent variables, and
- **R^2** represents the R-squared values determined by the model.

① k ↑, R^2 is nearly same
Adjusted R^2 will decrease

② k ↑, R^2 ↑
Adjusted R^2 will also increase

⇒ Ridge Regression

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n-1)}{(n-k-1)}]\}$$

So, if R-squared does not increase significantly on the addition of a new independent variable, then the value of Adjusted R-squared will actually decrease.

On the other hand, if on adding the new independent variable we see a significant increase in R-squared value, then the Adjusted R-squared value will also increase.