Module B: AI and Machine Learning

Lecture 21-22
Linear Regression

Kazi Saeed Alam, Assistant Professor, Dept. of CSE, KUET
email: saeed.alam@cse.kuet.ac.bd

Machine Learning

- ❖ Machine Learning is the science to make computers learn from data without explicitly program them and improve their learning over time in autonomous fashion.

- ❖ This learning comes by feeding them **data** in the form of observations and real-world interactions.”

- ❖ Machine Learning can also be defined as a tool to **predict** future events or values using past **data**.

Types of Data

- ***Based on Values***
  - *Continuous data (ex. Age – 0-100)*
  - *Categorical data (ex. Gender- Male/Female)*
- ***Based on pattern***
  - *Structured data (ex. Databases)*
  - *Unstructured data (ex. Audio, Video, Text)*

Types of Data- continued

- ***Labelled data*** *– consists of input output pair. For every set input features the output/response/label is present in dataset. (ex- labelled image as cat's or dog's photo)*

$$\{(x_1, y_1), (x_2, y_2), (x_3, y_3) \ldots\ldots\ldots\ldots\ldots (x_n, y_n)\}$$

- ***Unlabelled data****- There is no output/response/label for the input features in data. (ex. news articles, tweets, audio)*

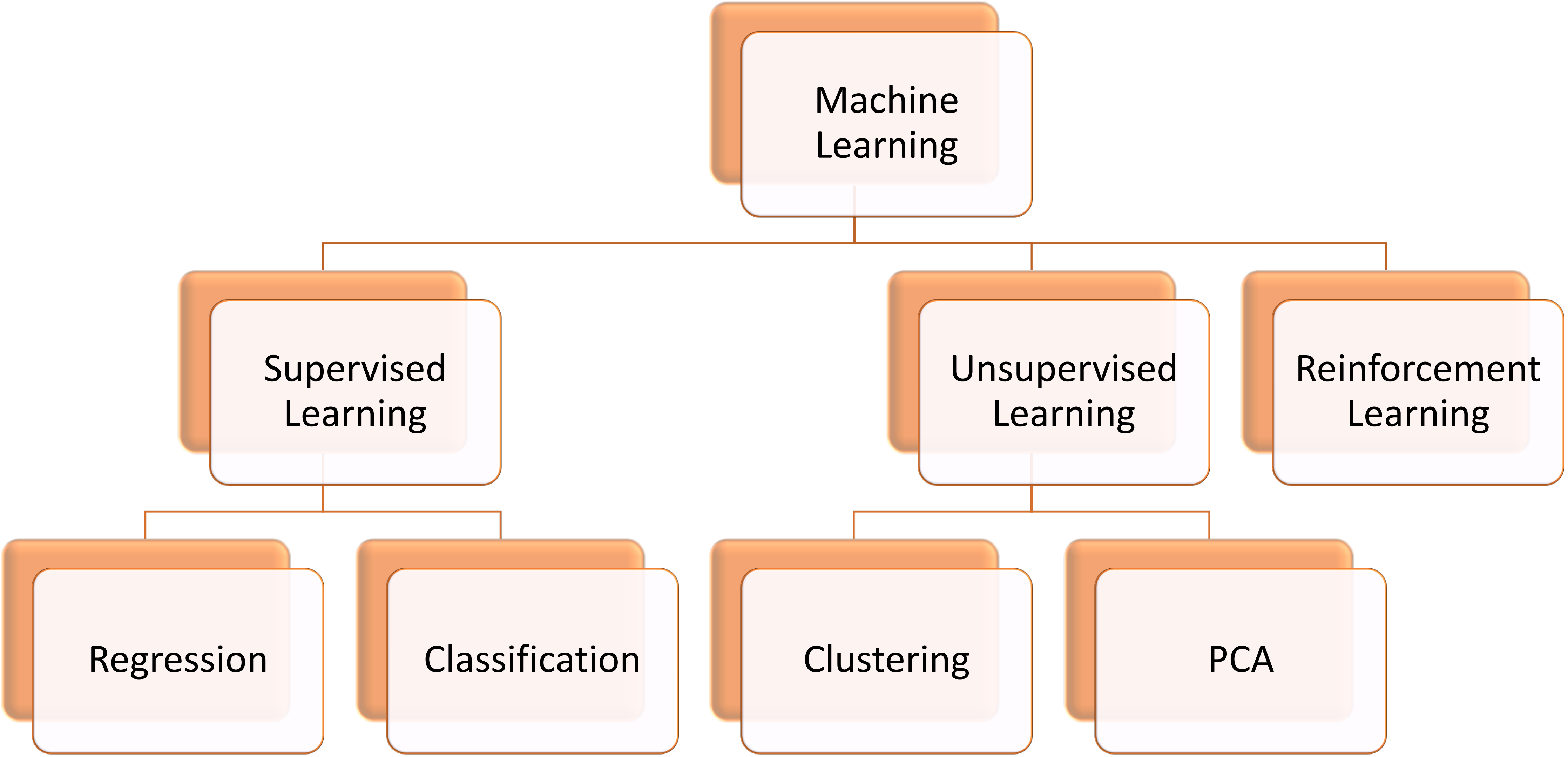
$$\{x_1, x_2, x_3 \ldots\ldots\ldots\ldots x_n\}$$

Types of Data- continued

- ***Training Data** – Sample data points which are used to train the machine learning model.*
- ***Test Data**- sample data points that are used to test the performance of machine learning model.*

Note- For modelling, the original dataset is partitioned into the ratio of 70:30 or 75:25 as training data and test data.

Types of Machine Learning



Supervised Learning

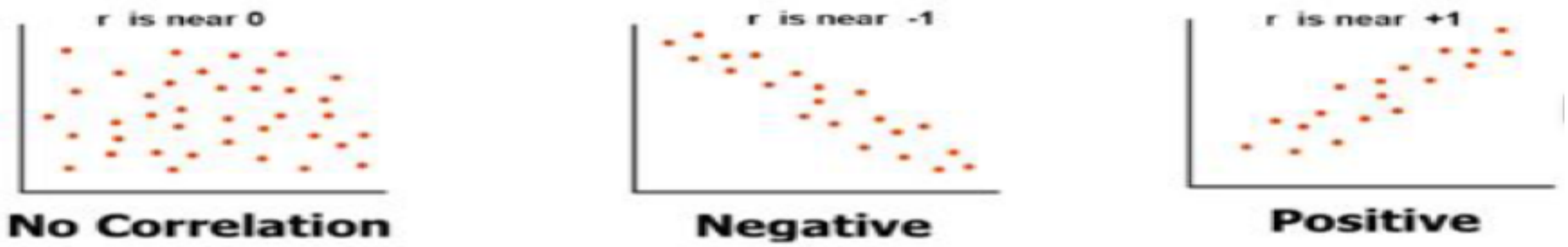
- Class of machine learning that work on externally supplied instances in form of predictor attributes and **associated target values**.

- The model learns from the training data using these ‘**target variables’** as reference variables.
  - Ex1 : *model to predict the resale value of a car based on its mileage, age, color etc.*

- The **target values** are the ‘correct answers’ for the predictor model which can either be a **regression model** or a **classification model.**

Motivation for learning

- It is being assumed that there exists a relationship/association between **input features** and **target variable.**
- Relationship can be observed by plotting a scatter plot between the two variables.



- Relationship measure can be quantified by calculating correlation between two the variables.

$$corr(x,y) = \frac{cov(x,y)}{var(x).var(y)} = \frac{\sum(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Linear Regression

- Linear regression is a way to identify a relationship between two or more variables and use these relationships to predict values of one variable for given value(s) of other variable(s).
- Linear regression assume the relationship between variables can be modelled through linear equation or an equation of line

$$\underbrace{y}_{\text{Dependent/Regressed variable}} = \underbrace{w_0}_{\text{Intercept}} + \overbrace{w_1}^{\text{Slope}} \underbrace{X}_{\text{Independent/Regressor variable}}$$

Multiple Regression

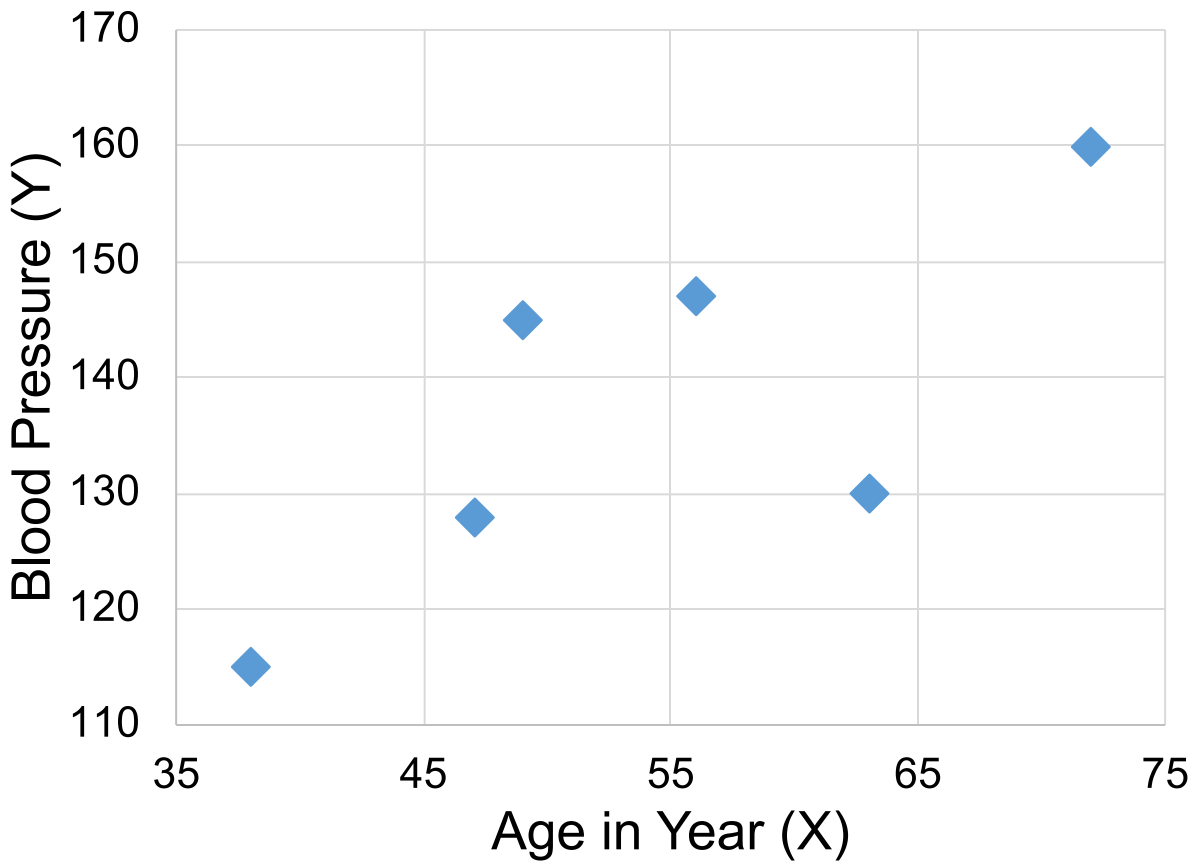
- Last slide showed the linear regression model with one independent and one dependent variable.
- In Real world a data point has various important attributes and they need to be catered to while developing a regression model. (Many independent variables and one dependent variable)

$$y = w_0 + w_1x_1 + w_2x_2 + w_3x_3 \ldots \ldots \ldots \mathrm{w_n x_n}$$

Regression –Problem Formulation

Let you have given with a data:

Age in Years **(X)**	Blood Pressure **(Y)**
56	147
49	145
72	160
38	115
63	130
47	128



Linear Regression

❖ For given example the Linear Regression is modeled as:

$$BloodPressure(y) = w_0 + w_1 AgeinYear(X)$$

OR

$$y = w_0 + w_1 X$$ – Equation of line

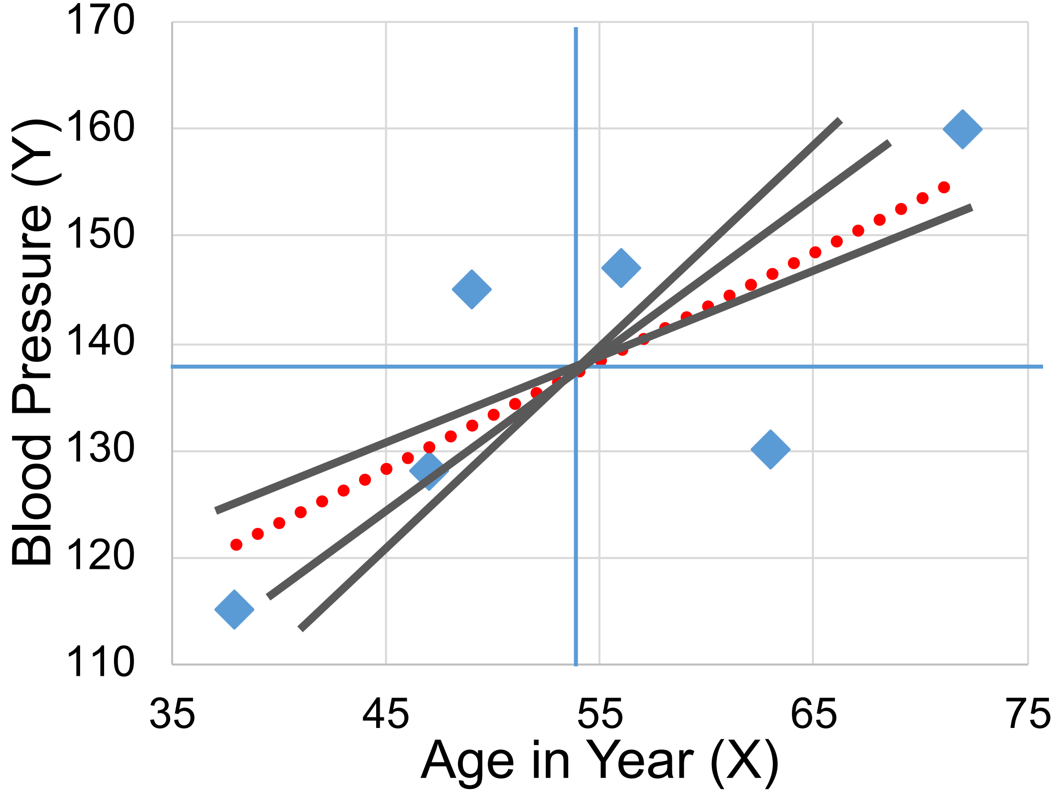
with w_0 $is\ intercept\ on\ Y\_axis\ and\ w_1\ is\ slope\ of\ line$

Blood Pressure - **Dependent Variable**

Age in Year - **Independent Variable**

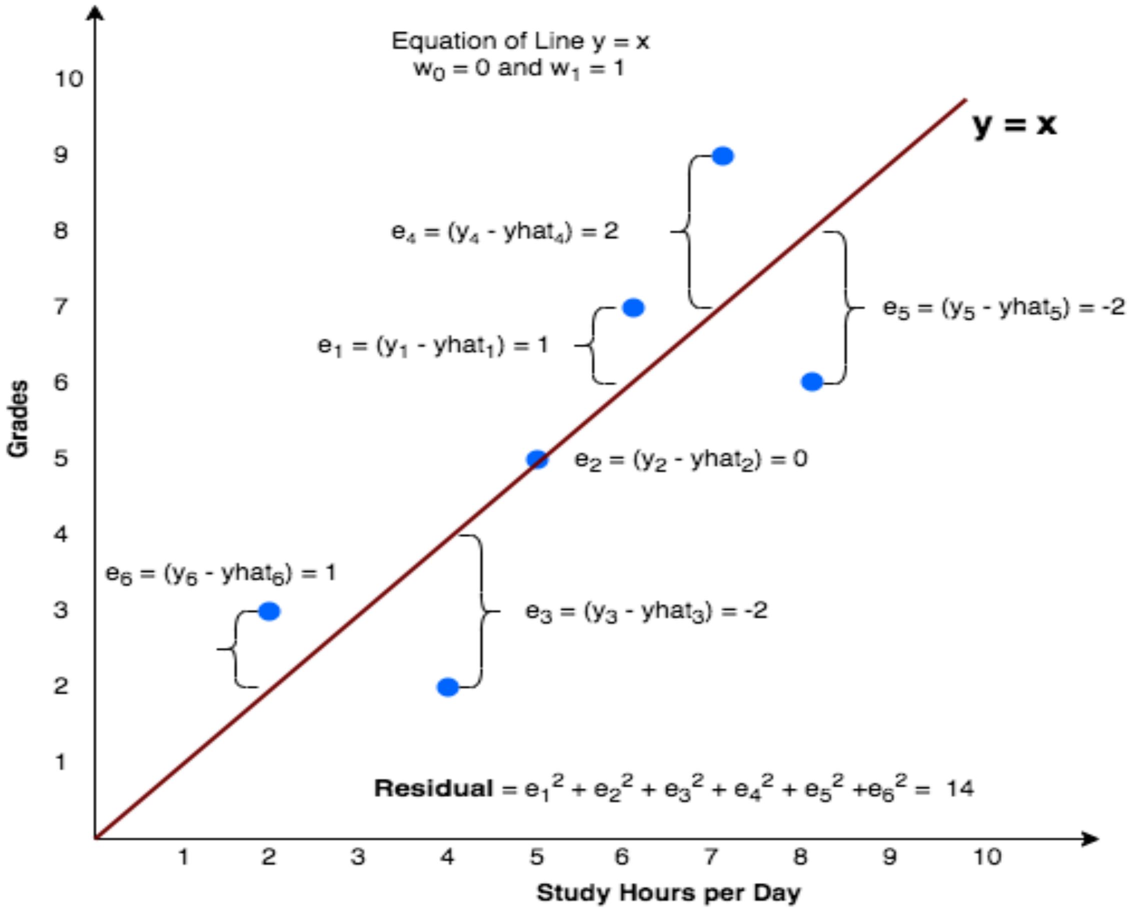
Linear Regression- Best Fit Line

- ❖ Regression uses line to show the trend of distribution.
- ❖ There can be many lines that try to fit the data points in scatter diagram
- ❖ The aim is to find **Best fit** Line



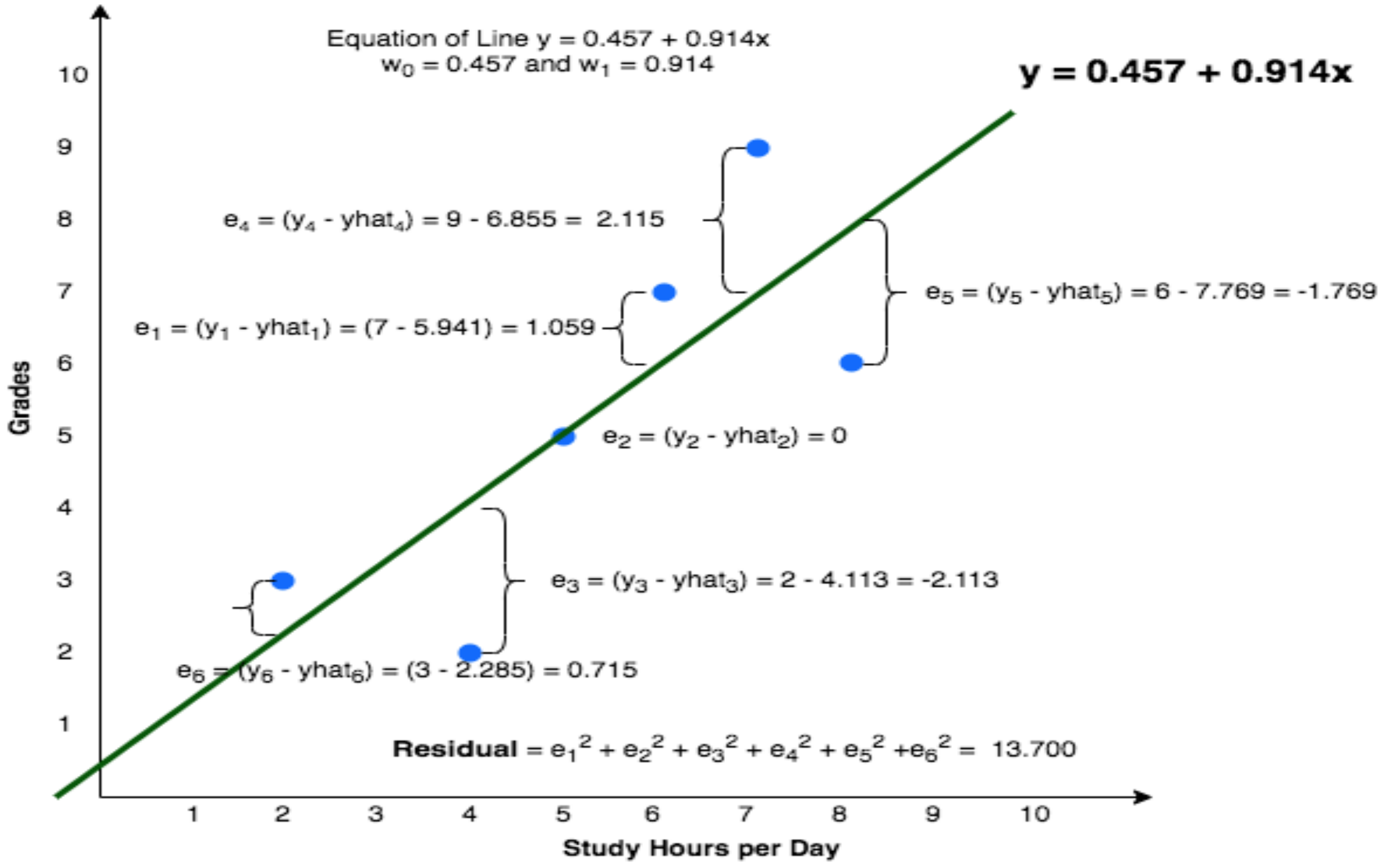
What is Best Fit Line

❖ Best fit line tries to explain the variance in given data. (minimize the total residual/error)



What is Best Fit Line

- Best fit line tries to explain the variance in given data. (minimize the total residual/error)



Linear Regression- Methods to Get Best

- ❖ Least Square
- ❖ Gradient Descent

Linear Regression- Least Square

Model: $Y = w_0 + w_1 X$

Task: Estimate $the\ value\ of\ w_0\ and\ w_1$

According to pr$inciple\ of\ least\ square\ the\ normal\ equations\ to\ solve\ for\ w_0\ and\ w_1$

$$\sum_{i=1}^{n} Y_i = n\, w_0 + w_1 \sum_{i=1}^{n} X_i \ldots\ldots\ldots\ldots\ldots (1)$$

$$\sum_{i=1}^{n} X_i Y_i = w_0 \sum_{i=1}^{n} X_i + w_1 \sum_{i=1}^{n} X_i^2 \ldots\ldots\ldots\ldots. (2)$$

Linear Regression–Least Square

Let divide the equation (1) by *n* (number of sample points) we get:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = w_0 + w_1 \frac{1}{n}\sum_{i=1}^{n} X_i$$

OR

$$\bar{y} = w_0 + w_1\bar{x} \ldots\ldots\ldots\ldots(3)$$

So line of regression will always passes through the points $(\bar{x}, \bar{y})$

Linear Regression–Least Square

Now we know :

$$cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y} \coloneqq \frac{1}{n}\sum_{i=1}^{n} x_i y_i = cov(x,y) + \bar{x}\bar{y} \text{ ....(4)}$$

and

$$var(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \quad \text{and} \quad var(y) = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \bar{y}^2$$

Dividing equation (2) by n and using equation (4) and (5) we get:

$$cov(x,y) + \bar{x}\bar{y} = w_0\bar{x} + w_1(var(x) + \bar{x}^2)\text{.........................(5)}$$

Linear Regression–Least Square

Now by using equation

$$\bar{y} = w_0 + w_1\bar{x}$$

and

$$cov(x, y) + \bar{x}\bar{y} = w_0\bar{x} + w_1(var(x) + \bar{x}^2)$$

We will get:

$$w_1 = \frac{cov(x, y)}{var(x)}$$

and

$$w_0 = \bar{y} - w_1\bar{x}$$

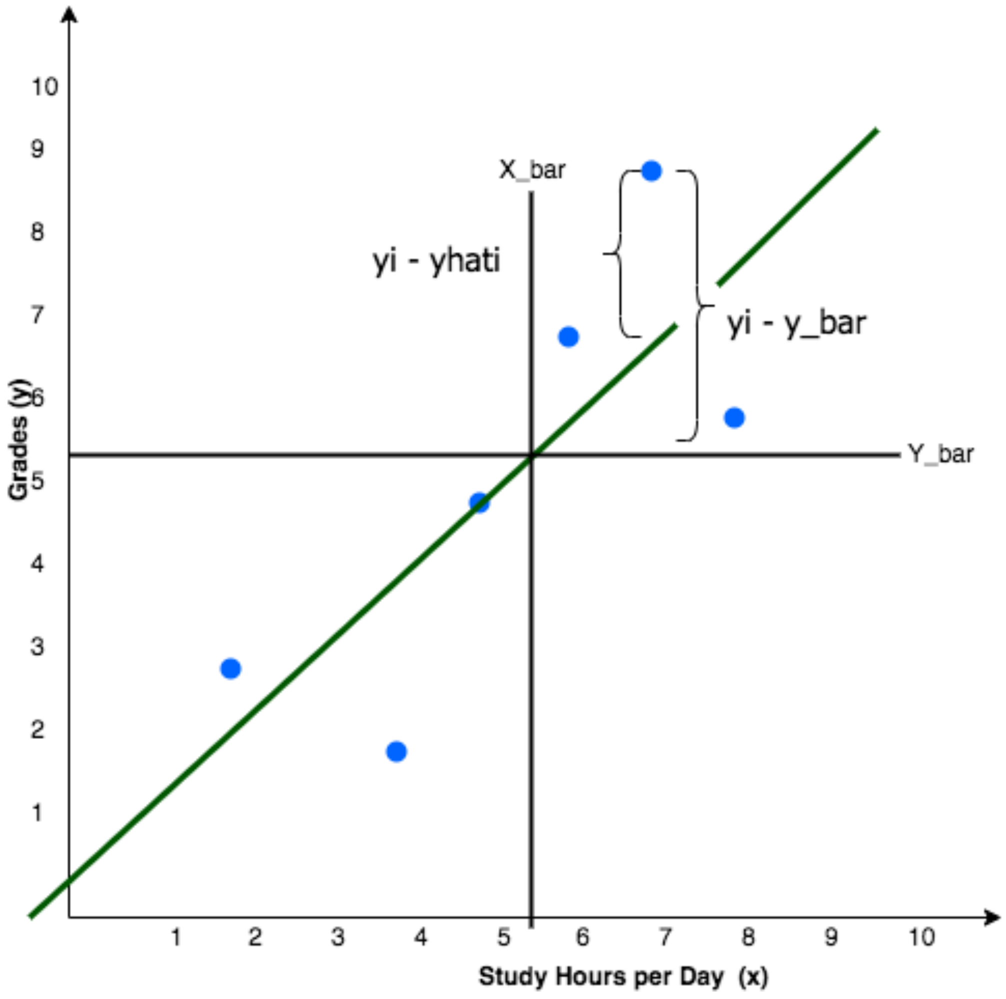
Performance metric for least square regression

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(y_i - yhat_i)^2}{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n-1)}{(n - k - 1)}$$

Interpretation:

- $R^2 \in [0,1]$: A value close to 1 indicates a good fit.
- $R^2 = 0$: The model explains none of the variability in y.
- A negative R^2 can occur if the model performs worse than a horizontal line at $\bar{y}$.



Linear Regression- Gradient Descent

Model: $Y = \ w_0 + w_1 X$

Task: Estimate $the\ value\ of\ w_0\ and\ w_1$

Define the cost function,

$$cost(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i \ - yhat_i)^2$$

Objective of gradient Descent

$$\min_{w_0, w_1} cost(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i \ - (w_0 + w_1 x_i))^2$$

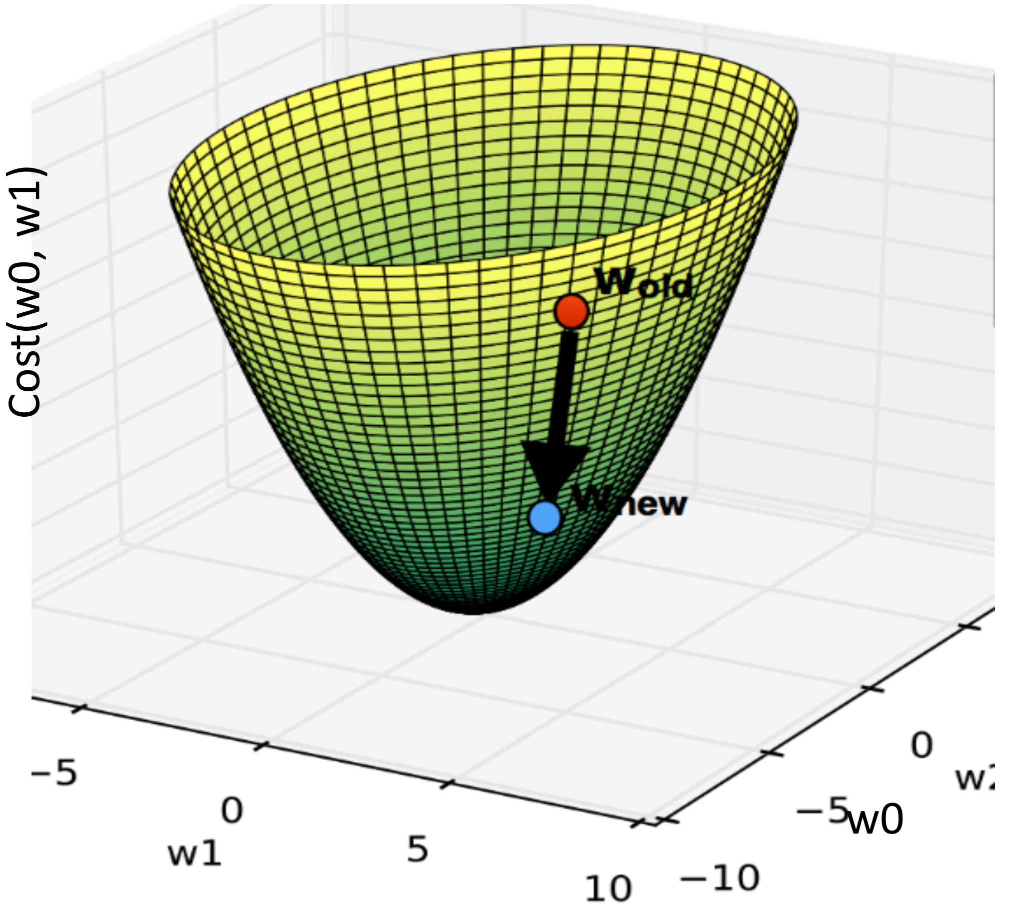
Linear Regression- Gradient Descent

Model: $Y = w_0 + w_1 X$

Task: Estimate $the\ value\ of\ w_0\ and\ w_1$

the objective,

$$\min_{w_0, w_1} cost(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$



Linear Regression- Gradient Descent

- Gradient descent works if following steps:
  1. Initialize the parameters to some random variable
  2. Calculate the gradient of cost function w. r. t. to parameters
  3. Update the parameters using gradient in opposite direction.
  4. Repeat step-2 and step-3 for some number of times or till it reaches to minimum cost value.

Linear Regression- Gradient Descent

$$cost(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))^2$$

Calculating gradients of cost function:

$$gradw_0 = \frac{\partial cost(w_0, w_1)}{\partial w_0} = \frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))(-1)$$

$$gradw_1 = \frac{\partial cost(w_0, w_1)}{\partial w_1} = \frac{2}{n}\sum_{i=1}^{n}(y_i - (w_0 + w_1 x_i))(-x)$$

Parameter update:

$$w_0 = w_0 - learningrate * gradw_0$$

$$w_1 = w_1 - learningrate * gradw_1$$

Performance metric for gradient based regression

Root Mean Square Error (**RMSE**) is the standard deviation of prediction errors.

$$RMSE = \sqrt{\frac{(y_i - yhat_i)^2}{n}}$$

Mean absolute error (MAE) is a measure of difference between two variables.

$$MAE = \frac{|y_i - yhat_i|}{n}$$