SOLENT UNIVERSITY SOUTHAMPTON

TEF 2023 Gold

INTRODUCTION TO AI (COM727)

(MSc Applied AI and Data Science)

Week 5

Dr Kashif Talpur
Lecturer Computing (AI & Data Science)

What is this week about?

- This week you will

  - focus on Linear Regression and its types.
  - understand the difference between regression and classification problems
  - learn the difference between simple, multiple linear and multivariate regression.
  - learn linear Regression Model – line fitting, loss function.
  - Learn about Gradient Descent and implement in Python
  - Explore concepts like learning rate, number of iterations, convergence, local and global minima
  - Spend most of the time working on practical activities.
  - Take a quiz at the end.

Regression and Classification Problems

SOLENT UNIVERSITY

- Supervised Machine Learning is applied to mainly two types of problems
- Regression and Classification

- **Classification Problems**
  - We predict the class label, the output that we want to predict falls in one of the known classes and we need to determine which class it falls in. e.g., given a flower picture find out which specie it belongs to.
  - **Binary Classification:** only two classes, (Yes/No or True/False) e.g., is a given transaction fraudulent or not, or is a given email junk or not?
  - **Multiclass Classification:** More than two classes, identify which specie a given animal picture belongs to
  - The output is categorical rather than numerical.

- **Regression Problems**
  - We predict a **continuous number** i.e., a real number and the set of possible values could be infinite. Examples: predict the share price of given stock tomorrow, or predict a person’s annual income using age.

Regression in Machine Learning

- **Regression Examples**
  - Understanding the relationship between
    - ✓ **Advertising Spending** and **Revenue**

      revenue = β0 + β1(ad spending)
    - ✓ **Drug Usage** and **Blood Pressure of Patients**

      blood pressure = β0 + β1(dosage)
    - ✓ **Fertilizer** and **Water** on **Crop Yields**

      crop yield = β0 + β1(amount of fertilizer) + β2(amount of water)
    - ✓ **Prescription** and **Player Performance**

      points scored = β0 + β1(yoga sessions) + β2(weightlifting sessions)

- **More Examples**
  - Predicting House Value

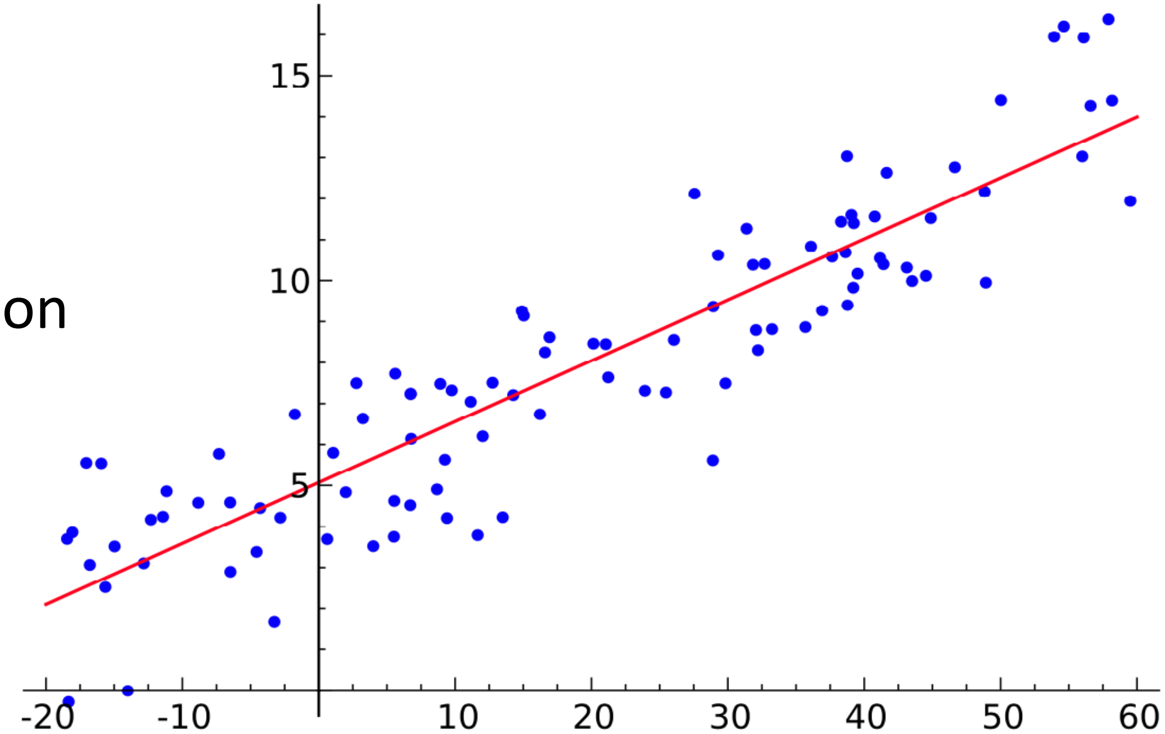
    Actual Price: £100,000
    Predicted 1: £99,950 (Very Good Prediction)
    Predicted 2: £50,000 (Very Bad Prediction)
  - Predicting Car Premium

    Using Location, Age, History etc.

Regression Techniques

SOLENT UNIVERSITY

- Linear Regression
  - Single Linear Regression
  - Multiple Linear Regression
  - Multivariate Linear Regression
- Ridge Regression
- Lasso Regression
- Logistic Regression
- And many more



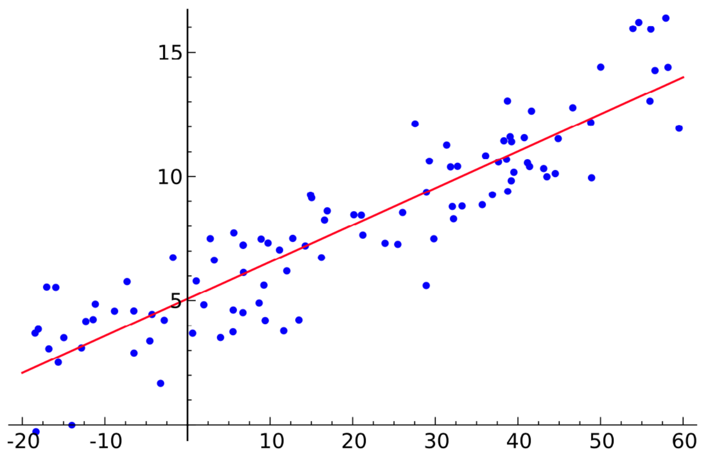
Single Linear Regression

- Its a branch of **statistics** that deals with **modelling the relationship** between **output (dependent variables) Y** and **input (independent variables) X**.

- It **assumes** there is **linear relationship** between **output** and **input** variables.

- Data is modelled using a **linear prediction function** and the **model parameters** are **estimated** from the data such that the model fits the data – such models are called **linear models**.

  - Given (x_i, y_i) for $i \in (1, n)$, predict y_k for $x_k \notin x_i$ for $i \in (1, n)$



Single Linear Regression

SOLENT UNIVERSITY

- Linear regression components:

$$Y = a + bX$$

b is the slope of regression line
a is the y intercept of regression line

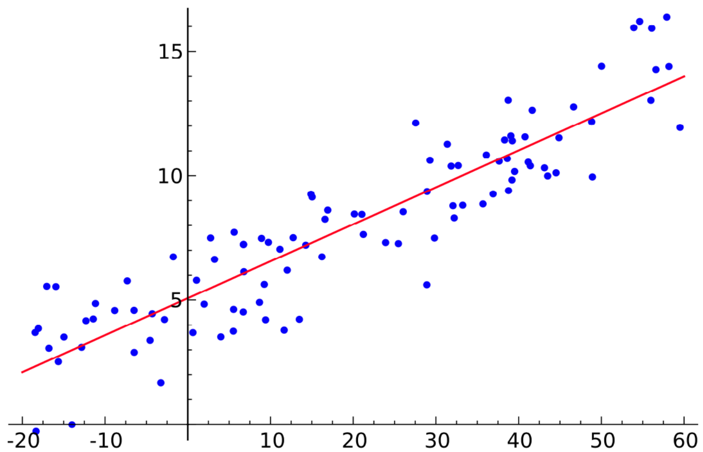
$$\text{where } b = r\frac{SD_y}{SD_x}$$

r is correlation coefficient: a statistic used to describe the strength of the relationship between two variables.
SD_X, SD_Y are the standard deviation of X and Y

$$a = Y' - bX'$$

X' and Y' is the mean of X and Y respectively

$$\text{slope} = \frac{N\sum XY - (\sum X)(\sum Y)}{N\sum X^2 - (\sum X)^2}$$



Single Linear Regression

- Can you predict Y for $X = 64$, using Linear Regression model for the following data?

	X	Y
x_1	60	3.1
x_2	61	3.6
x_3	62	3.8
x_4	63	4
x_5	65	4.1
x_6	64	?

Single Linear Regression

- Can you predict Y for $X = 64$, using Linear Regression model for the following data?

Solution:

Step 1: Total instance, N=5
Step 2: Calculate slop first
Step 3: Calculate intercept
Step 4: Calculate $y = a + bx$

	X	Y	X^2	$X \times Y$
x_1	60	3.1	3600	186
x_2	61	3.6	3721	219.6
x_3	62	3.8	3844	235.6
x_4	63	4	3969	252
x_5	65	4.1	4225	266.5
Total	311	18.6	19359	1159.7
x_6	64	?		

Single Linear Regression

SOLENT UNIVERSITY

- Go to your Google Collab. and code this:

```
import numpy as np
from sklearn.linear_model import LinearRegression
X = np.array([[1, 1], [1, 2], [2, 2], [2, 3]])
# y = 1 * x_0 + 2 * x_1 + 3
y = np.dot(X, np.array([1, 2])) + 3
reg = LinearRegression().fit(X, y)
reg.score(X, y)
```

- Download ".ipynb" file from the following link:
- https://scikit-learn.org/stable/auto_examples/linear_model/plot_ols.html#sphx-glr-auto-examples-linear-model-plot-ols-py
- On your Google Collab., upload the file and run it; observe the results.

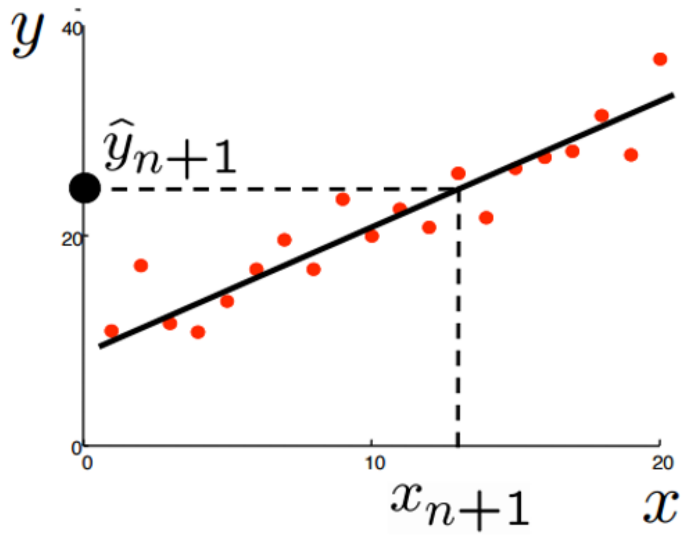
Multiple Linear Regression

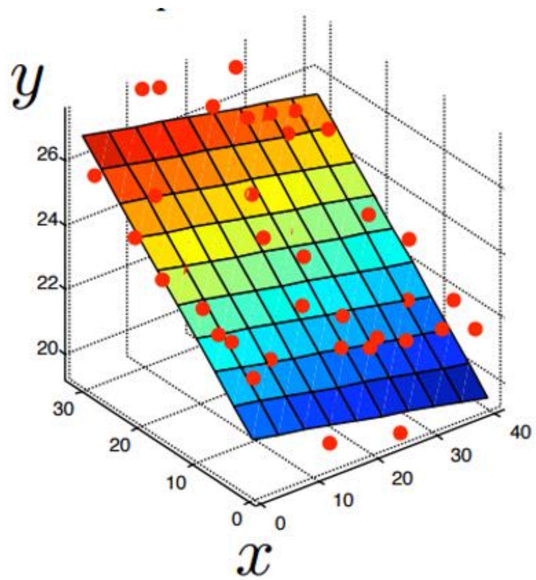
- More than one independent variables predict one dependent variable
- Predict the **price** of a house based on **square feet** and **number of bedrooms**.

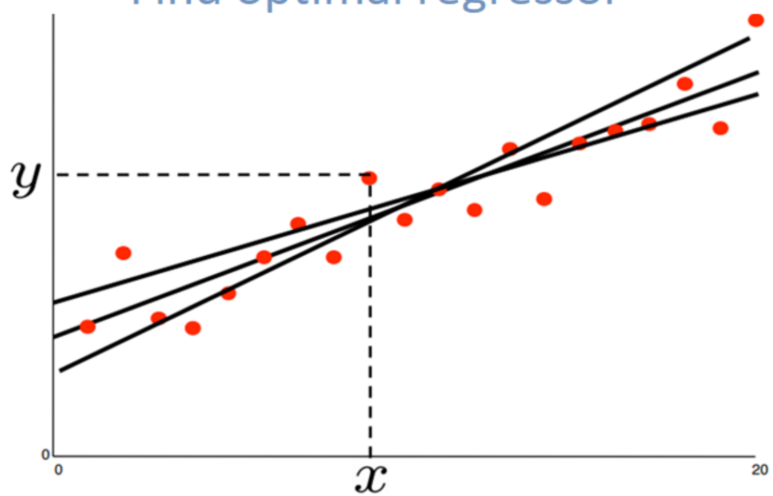
Estimate y' by a linear function of x:

$$y' = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$
$$y' = w^T x$$

w is the parameter to estimate







Multiple Linear Regression

SOLENT UNIVERSITY

- Least Mean Square (LMS) Algorithm

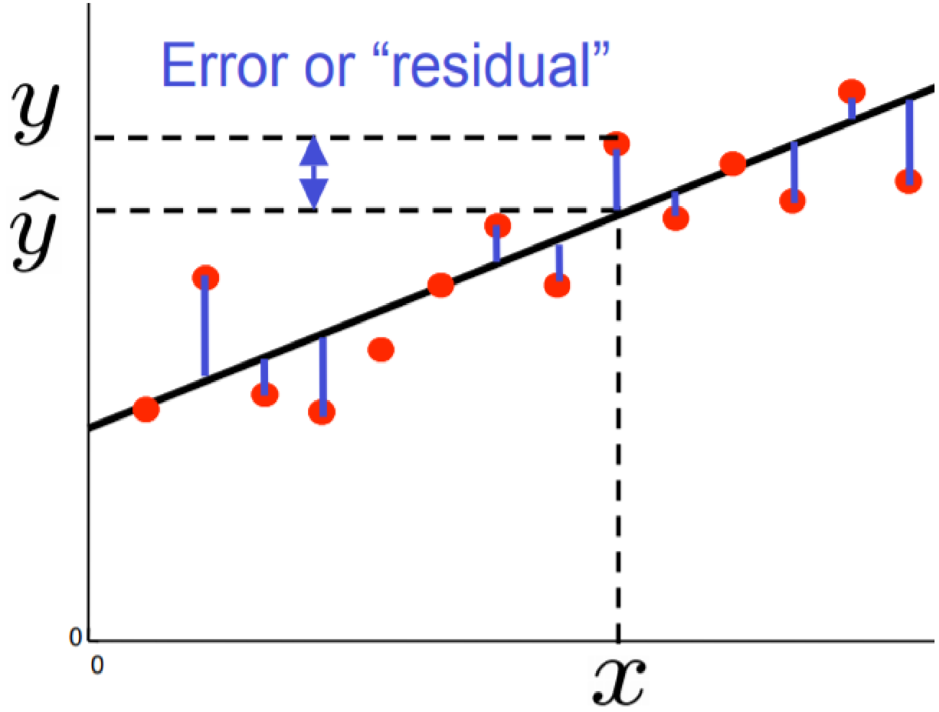
$$predicted_i = y_i{}' = w^T x_i$$

$$error_i = E_i = \frac{1}{2}(w^T x_i - y_i)^2$$

$$cost = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$$

$$E = \frac{1}{2}\sum_{i=1}^{n}(w^T x_i - y_i)^2$$

$$E = \sum_{i=1}^{n} E_i$$

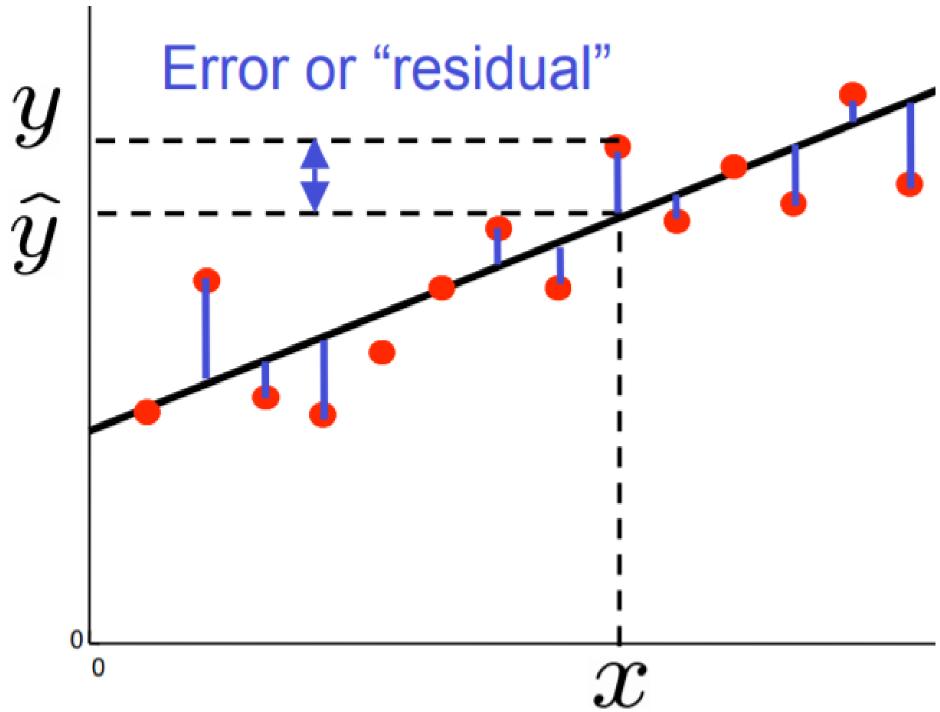


The objective is to optimize w: E is minimum

Multiple Linear Regression

- Least Mean Square (LMS) Algorithm
- Evaluation measure is Mean Squared Error (MSE)

Actual (Y)	Predicted (Y')	$Y' - Y$	$(Y' - Y)^2$
41	43.6	2.6	6.76
45	44.4	-0.6	0.36
49	45.2	-3.8	14.44
47	46	-1	1
44	46.8	2.8	7.84



Sum of Squared Error = 30.4

Mean Squared Error = $\frac{30.4}{5} = 6.08$

Multiple Linear Regression

- Least Mean Square (LMS) by ***Gradient Descent Algorithm***
- It is parameter estimation algorithm

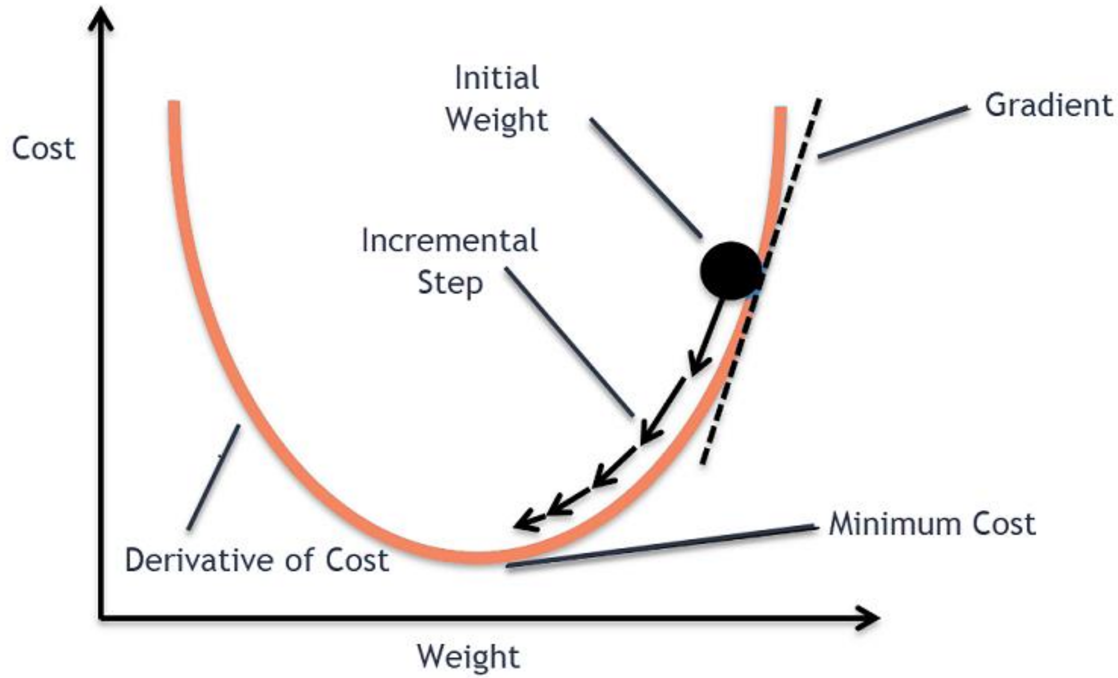
$$w^{t+1}: w - \alpha \frac{\partial}{\partial w} E$$

$$\frac{\partial}{\partial w} E = \sum_{i=1}^{n} \frac{\partial}{\partial w} E_i$$

$$\frac{\partial}{\partial w} E_i = \frac{1}{2} (w^T x_i - y_i)^2$$

$$\frac{\partial}{\partial w} E_i = (w^T x_i - y_i) \times x_i$$

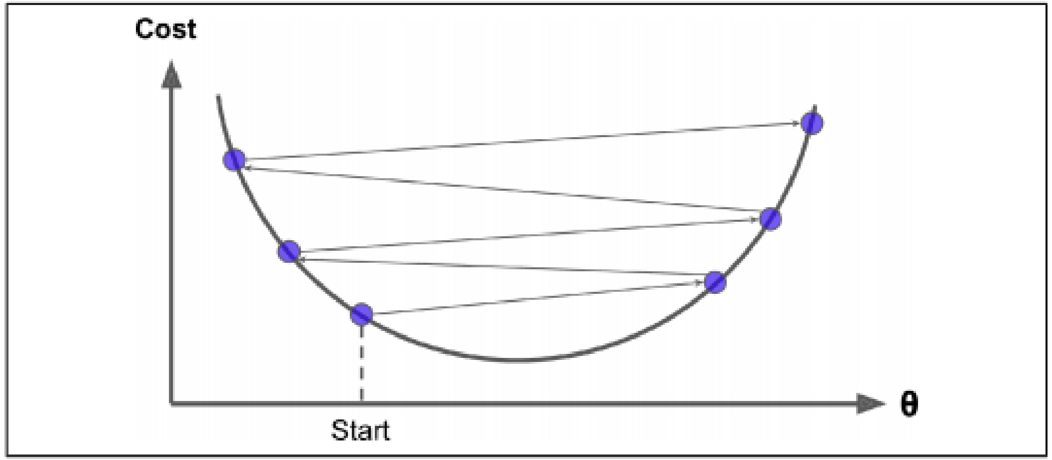
$$w_i^{t+1} = w_i^t - \alpha (w^T x_i - y_i) \times x_i$$

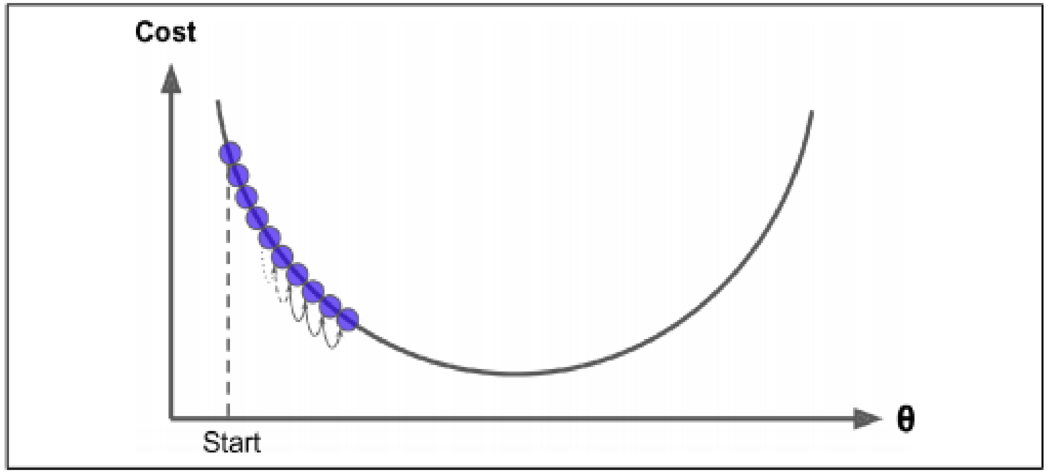


**Watch this video for easy understanding of the algorithm*
https://www.youtube.com/watch?v=sDv4f4s2SB8

Multiple Linear Regression

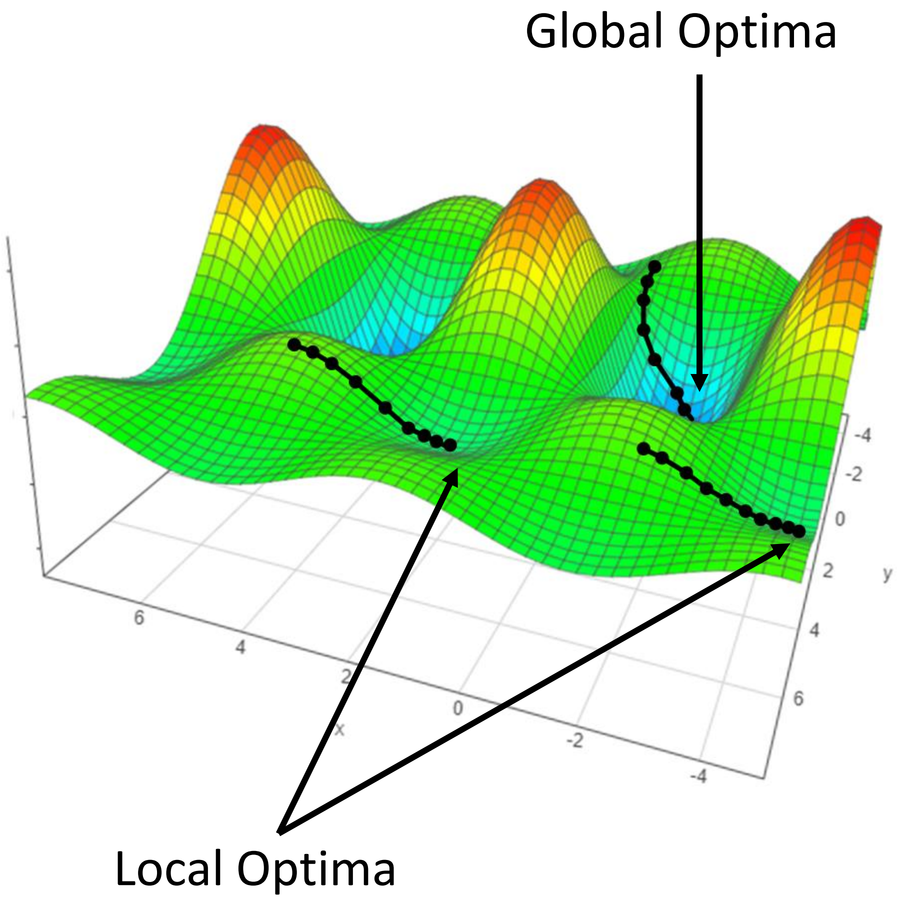
- Gradient Descent Algorithm **Convergence**
- Gradient Descent algorithm will **keep on tweaking values of parameters (W)** for each iteration and hope the algorithm will stop. But, **how do we know** when the algorithm has learned enough?
- We need to define **convergence**. Convergence is when the **loss stops changing** (or **changes very little**). At this time, we **hope** that the algorithm has found the best values.
- **Learning Rate** is important, as it determines the **step size by which we tweak W**
  - **Too high learning rate**, big jumps, chances of **missing minimum loss point**
  - **Too low learning rate**, tiny steps, **too many iterations**, high chances of finding minimum loss point

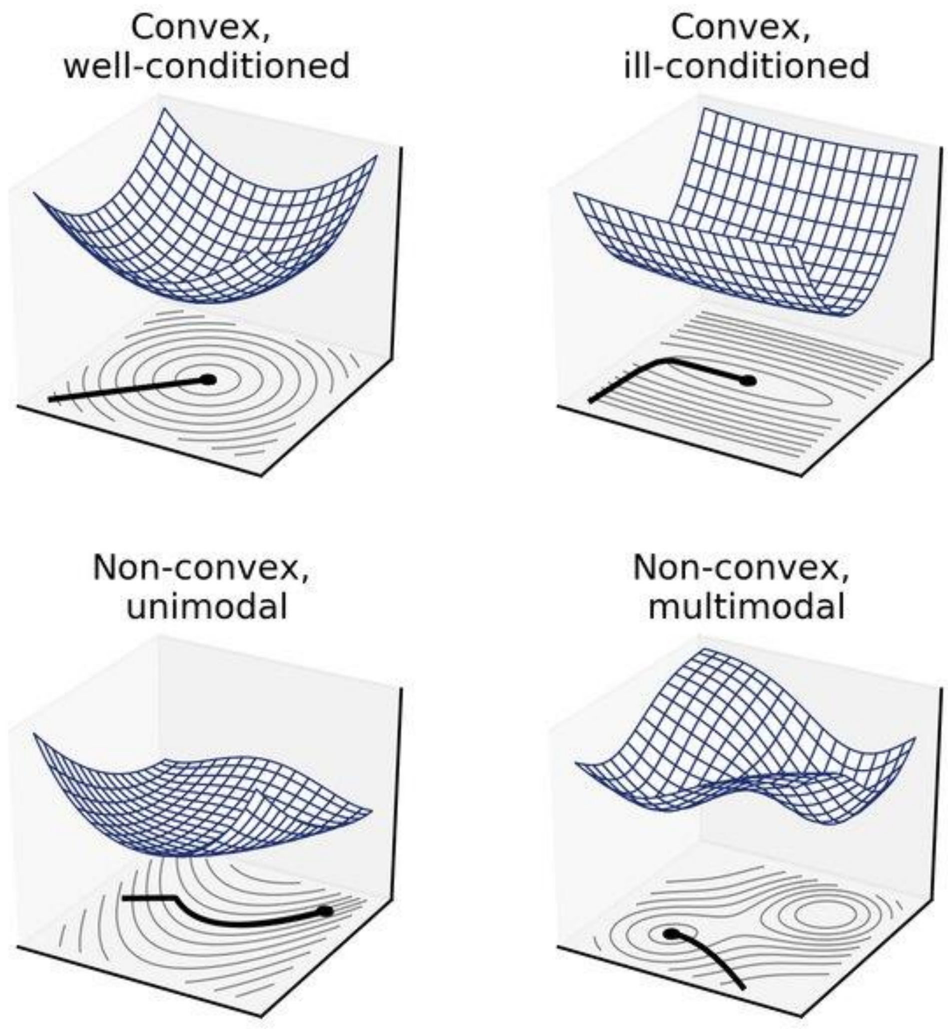




Multiple Linear Regression

- Types of optimization problem landscapes





Multivariate Linear Regression

- Also known as **multivariable regression**
- It deals with **multiple independent variables** and **multiple dependent variables**.
- It involves modeling the **relationship** between multiple independent variables (**X1, X2, X3, etc.**) and multiple dependent variables (**Y1, Y2, Y3, etc.**).
- Here, the model includes **multiple equations**, each of which **models** the **relationship** between **a specific set of independent and dependent variables**.
- This technique is used when you want to **understand** how **multiple independent variables collectively affect multiple dependent variables**. It's common in fields like **economics, social sciences, and engineering**.

Multivariate Linear Regression

SOLENT UNIVERSITY

- One common application of multivariate regression is in the **field of finance**, specifically in **portfolio management**. Here's a simplified example:

- **Problem:** An investment analyst wants to understand how various **economic factors** affect the **performance of a portfolio** consisting of different types of assets (stocks, bonds, and real estate). The analyst has collected historical data for the following variables:

- The goal is to build a multivariate regression model to understand how changes in these economic factors (independent variables) influence the performance of the investment portfolio (dependent variables). Following is the model:

$$Return = w_0^1 + w_1^1.GDP + w_2^1.Infl + w_3^1.IntRate + w_4^1.Unemp + \epsilon_1$$

$$Risk = w_0^2 + w_1^2.GD + w_2^2.Infl + w_3^2.IntRate + w_4^2.Unemp + \epsilon_2$$

- Download the file *MultivariateRegression.ipynb* from SOL; upload on your Google Collab.; run the code and observe results.

Let's upload today's work on your GitHub