Welcome to Data Science Online Bootcamp

Week#3_Day#4

dφ

Democratizing Data Science Learning

Learning Objectives

Statistics Refresher

Cost Function

Gradient Descent

Statistics Refresher

Measures of Central Tendencies

- **Mean:** The mean is the average of a data set. For example, take a list of numbers: 10, 20, 40, 10, 70
  Mean = (10 + 20 + 40 + 10 + 70) / 5 = 30

- **Median:** The median is the middle of the set of numbers.
  To find the median, first we sort the list of numbers:
  10, 10, 20, 40, 70
  The exact middle number i.e. 20 is the median.

- **Mode:** The mode is the most common number in a data set.
  In above list of numbers, 10 has occurred 2 times while other three numbers are occurred one time each.
  So, the mode is 10 here.

Measures of Dispersion

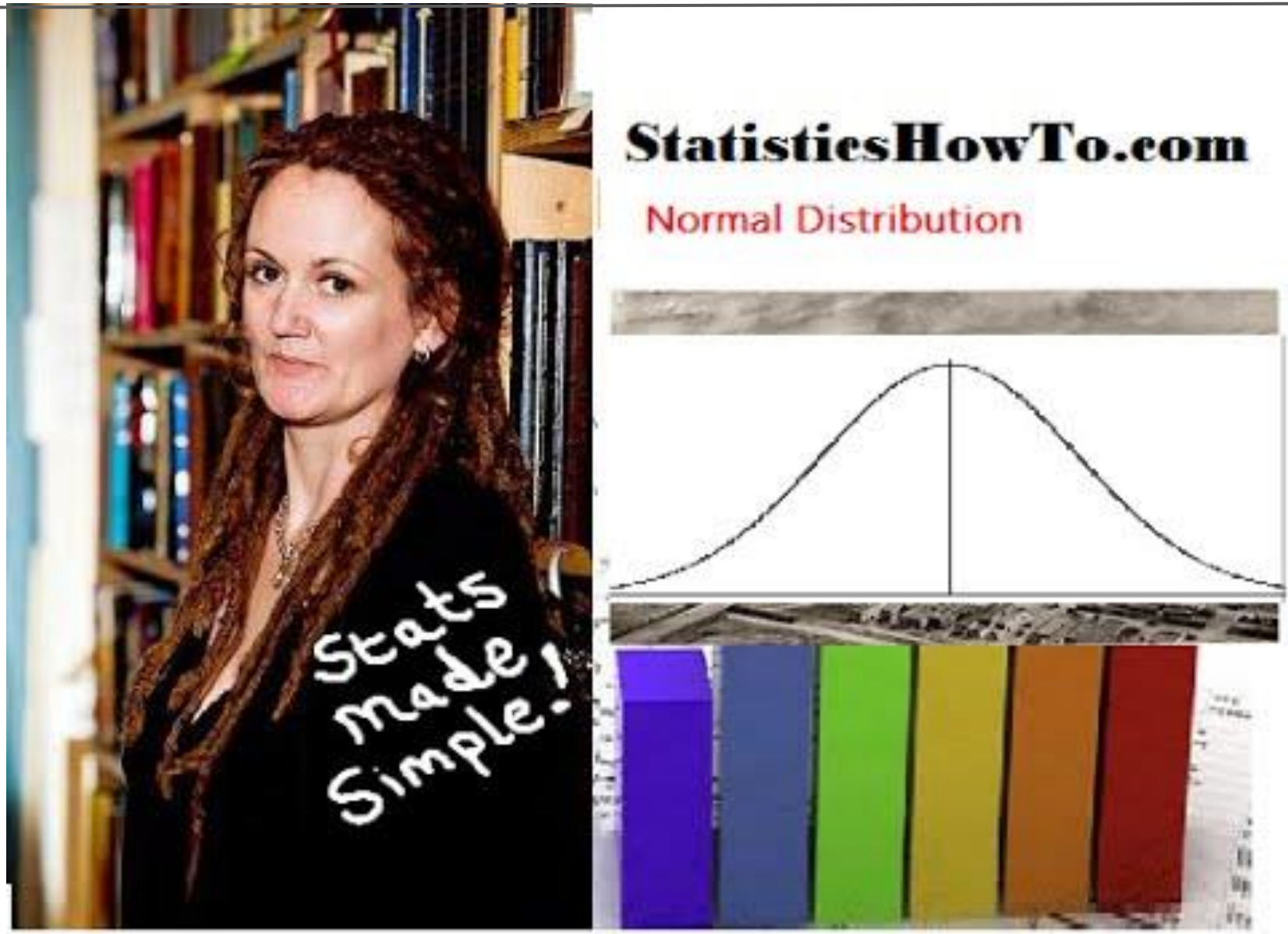
- **Range:** It is the difference between highest value and the lowest value in the data set.
  For a given list of numbers: 10, 20, 40, 10, 70 the range is 70 - 10 = 60.

- **Variance:** The average of the squared differences from the mean.
  Steps to calculate variance:
  - Calculate mean (mean is nothing but average)
  - Find difference of each data from mean
  - Square all the differences
  - Take the average of the squares.

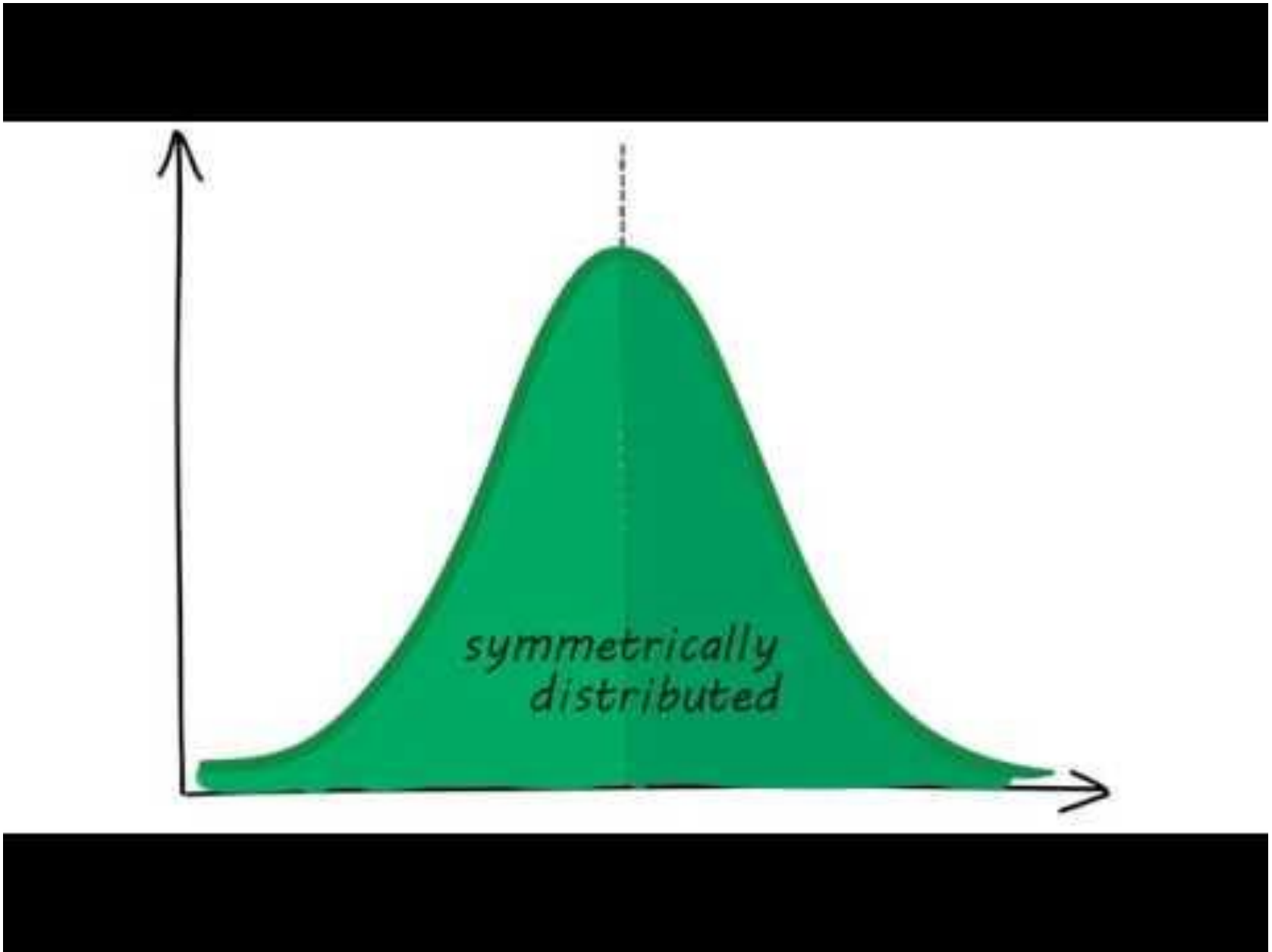
- **Standard Deviation:** It shows you how much your data is spread out around the mean. Its symbol is σ (the greek letter sigma). **It is the square root of the variance.**

$$Standard\ Deviation = \sqrt{Variance}$$

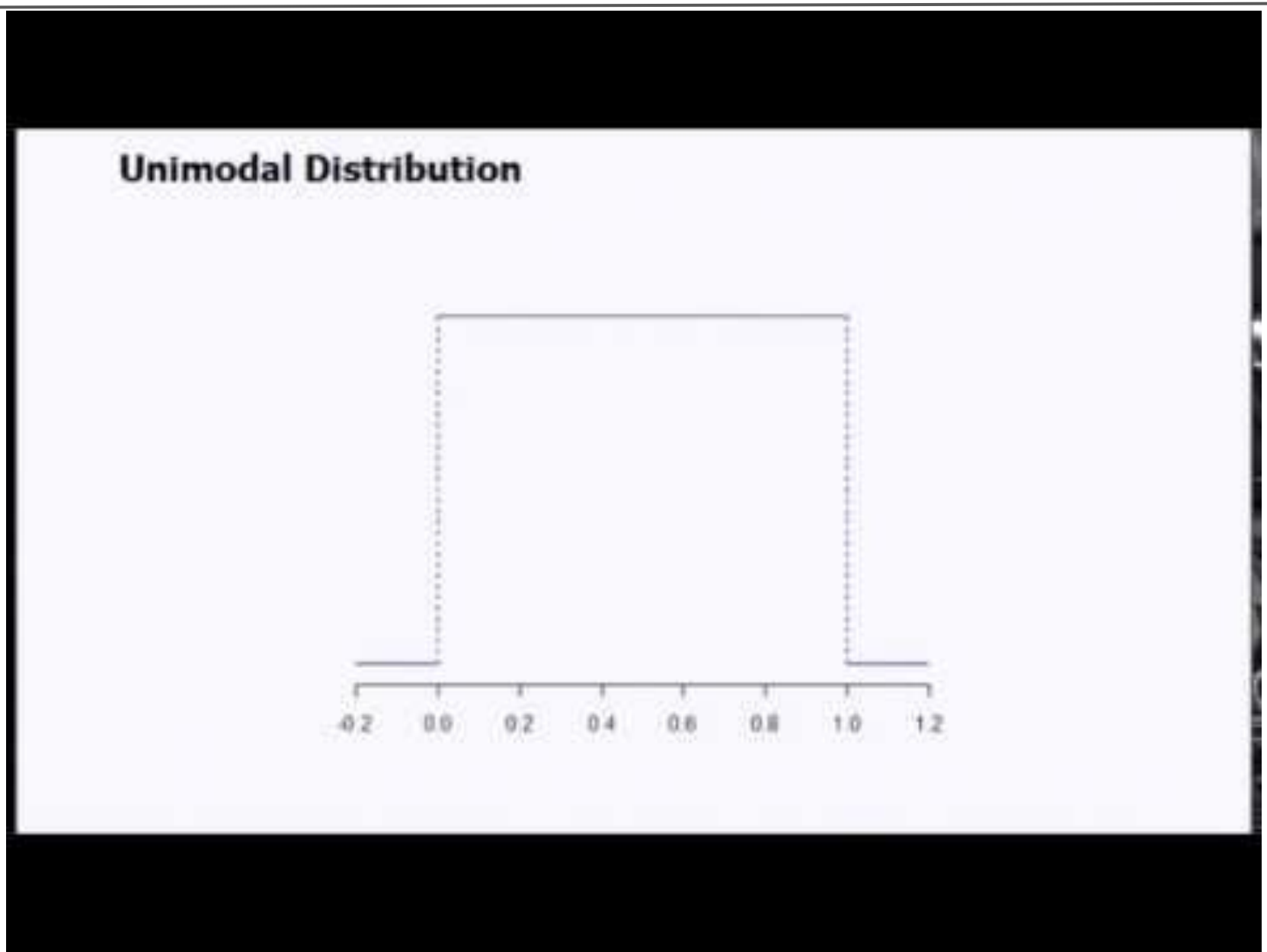
Normal Distribution



Normal Distribution in 60 secs



Unimodal and Multimodal Distribution



Now coming back to Linear Regression

This will be in-continuation of our yesterday's slides!

Cost Function

What?

Now that we built a model, we need to measure its performance right? and understand if it works well or not. Cost function measures the performance of a Machine Learning model for given data. It quantifies the error between predicted values and expected values and presents it in the form of a single real number.

Depending on the problem Cost Function can be formed in many different ways. The purpose of Cost Function is to be either:

- **Minimized** - then returned value is usually called **cost, loss or error.** The goal is to find the values of model parameters for which Cost Function return as small number as possible.
- **Maximized** - then the value it yields is named a **reward**. The goal is to find values of model parameters for which returned number is as large as possible.

What is predicted and expected value?

- **Predicted value:** As the name says is the predicted value of your machine learning model.

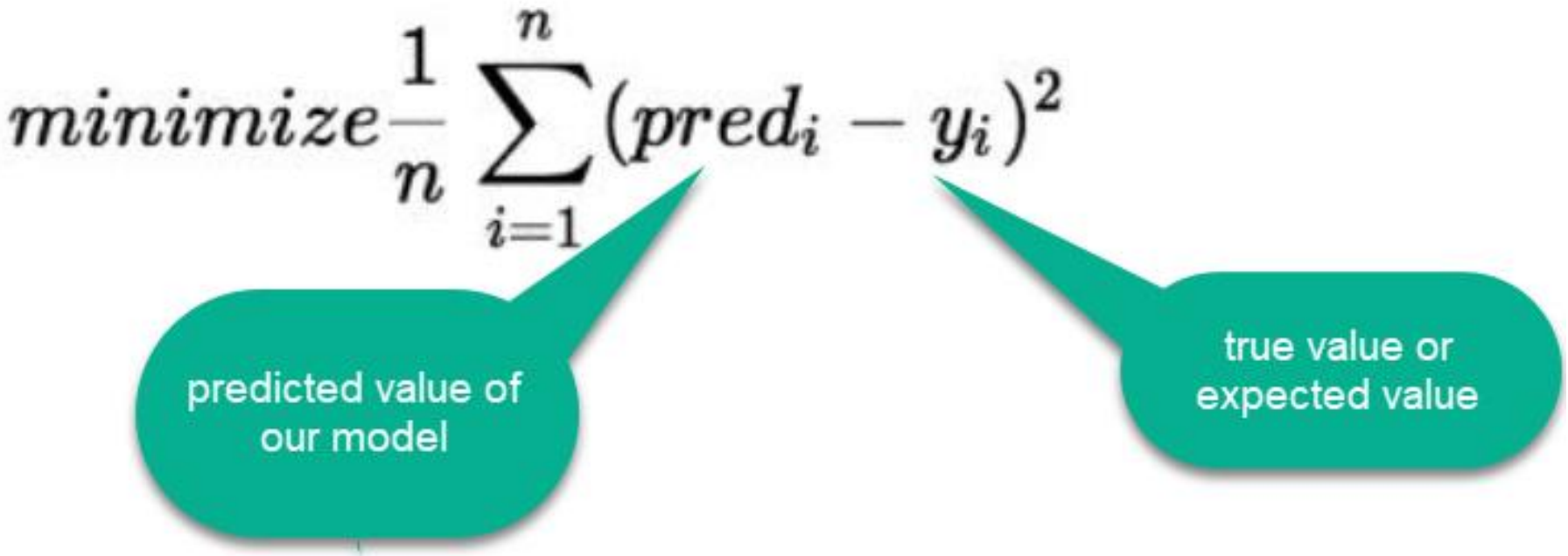
- **Expected value:** Is the true value(or the label present in your data)

Often machine learning models are not 100% accurate or perfect, then tend to deviate from the true value or expected value.

Explaining with an example: If we are predicting the age of a person based on few input variables or features.

- Our machine learning model predicted the age as 28 years
- However, the actual age of the person is 29 years.
- Here **28 years is predicted value** and **29 years is expected value or true value.** As data scientists, we try to minimize the error while building models.

Cost Function



The difference between the true value and the model's predicted value is called **residual.**

Cost Function Types/ Evaluation Metrics

There are three primary metrics used to evaluate linear models (to find how well a model is performing):

1. Mean Squared Error:
2. Root Mean Squared Error
3. Mean Absolute Error

Mean Squared Error (MSE)

- MSE is simply the average of the squared difference between the true target value and the value predicted by the regression model.

- As it squares the differences, it **penalizes** (gives some penalty or weight for deviating from the objective) **even a small error** which leads to **over-estimation of how bad the model is.**

$$MSE = \frac{1}{n}\Sigma\underbrace{\left(y - \hat{y}\right)^2}_{\text{The square of the difference between actual and predicted}}$$

Root Mean Squared Error (RMSE)

- It is just the square root of the mean square error.

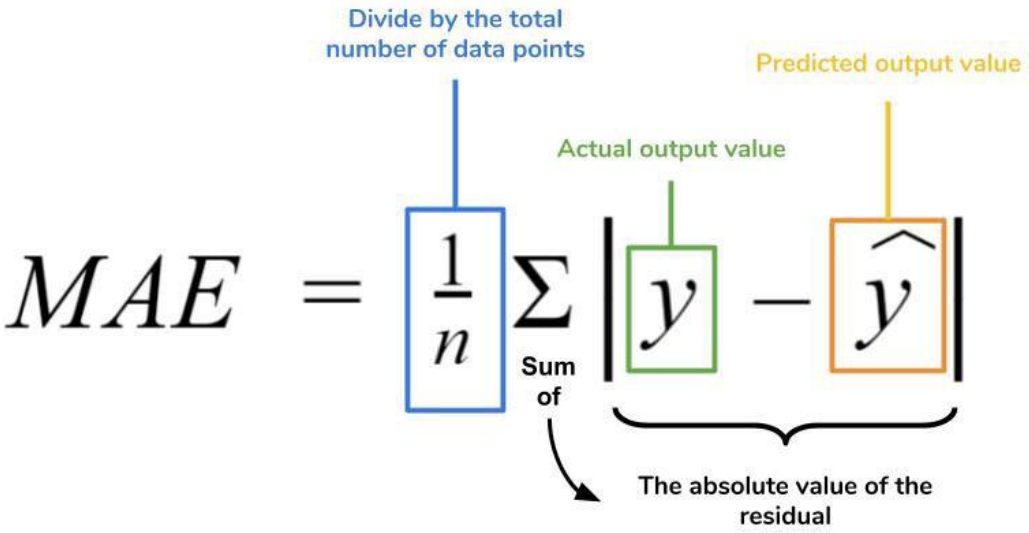
- It is preferred more in some cases because the errors are first squared before averaging which poses a high penalty on large errors. This implies that **RMSE is useful when large errors are undesired.**

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Mean Absolute Error(MAE)

- MAE is the absolute difference between the target value and the value predicted by the model.

- MAE **does not penalize the errors as effectively as mse** making it not suitable for use-cases where you want to pay more attention to the outliers.



Gradient Descent

As Data Scientists, we always want to optimise our algorithms and go for the best ones. Gradient Descent comes is one of those that helps us do this!

Gradient Descent is an optimization technique in the machine learning process which minimizes the cost function. Every machine learning algorithm has a cost function.

For now, we are not getting too much into how it works. We will learn about it as we proceed. You reading this article for further reading:

https://mc.ai/a-dummies-guide-to-gradient-descent-and-backpropagation/

Recap

- ➔ Linear regression is used to predict a value (like the sale price of a house).
- ➔ Given a set of data, first try to fit a line to it.
- ➔ The cost function tells you how good your line is.
- ➔ You can use gradient descent to find the best line.

Slides Download Link

https://docs.google.com/presentation/d/17g_wfuRvf8jm1w8zk8XR0LzoNypz15aPPYPedepqmPM/edit#slide=id.g89ba29404f_0_12

That’s it for the day. Thank you!

Feel free to post any queries in the #help channel on Slack

Optional Reading Material

If you are inquisitive to learn more about gradient descent refer the below video:

- https://www.youtube.com/watch?v=sDv4f4s2SB8
- https://youtu.be/vsWrXfO3wWw