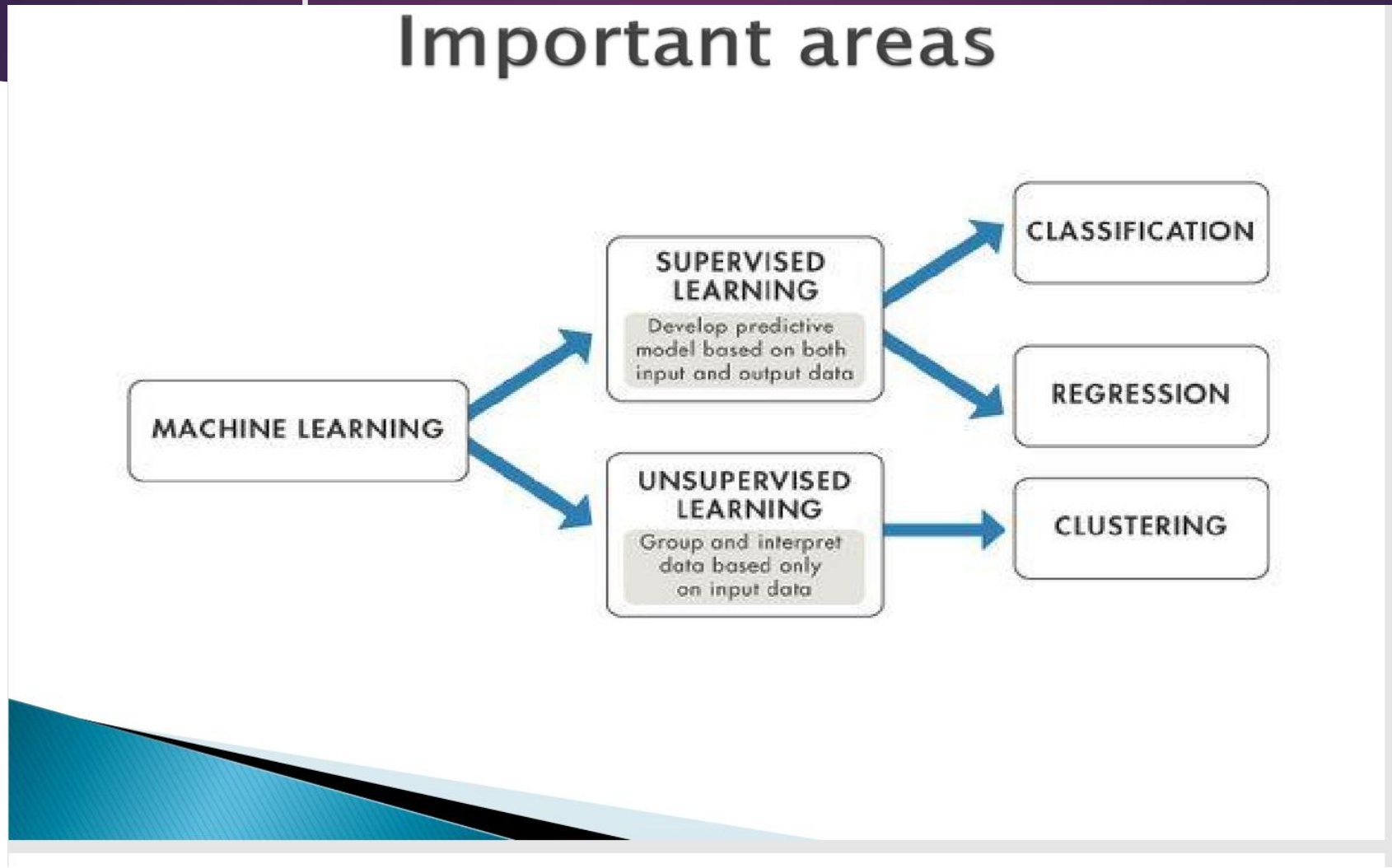
LINEAR REGRESSION

MACHINE LEARNING SESSION 2

Road Map


Important areas
MACHINE LEARNING
SUPERVISED LEARNING
Develop predictive model based on both input and output data
UNSUPERVISED LEARNING
Group and interpret data based only on input data
CLASSIFICATION
REGRESSION
CLUSTERING

INTRODUCTION

- Machine learning, more specifically the field of predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible

- As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

INTRODUCTION

- Linear regression is a **Supervised linear model** , e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).
- When there is a single input variable (x), the method is referred to as **simple or univariate linear regression** . When there are multiple input variables, literature from statistics often refers to the method as **multiple or multivariate linear regression**.

NOTATIONS

- m= number of training examples
- **x**'s = "input" variable / features
- y's = "output" variable /"target" variable

Hypothesis (Z)-for univariate LR

$$Z_{(W,b)}(x) = \mathrm{b} + \mathrm{Wx}$$

Parameters:- W, b(usually called bias)

Examples:

- Why bias?

  (Refer class notes)

Idea

- Idea of training the algorithm is :

  choose W , b so that Z(x) (predicted) is close to ‘y’ (actual value)

  for all the training examples (x,y)

- Hence some sort of cost function is required ,

  Squared Error Function is used

COST FUNCTION

- $J(W,b) = \frac{1}{2m}\sum_{i=1}^{m}(Z\left(x^{(i)}\right) - y^{(i)})^2$

- According to the proposed idea ,

$$\underset{(W\ ,\ b)}{\boldsymbol{minimize}}\ \boldsymbol{J(W,b)}$$

This is called Objective function of Linear Regression

Vectorization

- Given hypothesis is of the form,

$$Z_{(W,b)}(x) = \mathrm{b} + \mathrm{Wx}$$

▶ Why vectorization?

Python 'for' loops are inherently slower than their C counterpart. This is why 'numpy' offers vectorized actions on 'numpy' arrays. It pushes the 'for' loop you would usually do in Python down to the C level, which is much faster. 'numpy' offers vectorized ("C level for loop") alternatives to things that otherwise would need to be done in an element-wise manner ("Python level for loop).

Vectorized Hypothesis

- On using vectorisation, the new hypothesis become the ‘dot’ product of param(array of parameters) and X(array of input training examples)

$$Z_{(W,b)}(x) = \mathrm{X} \,.\, param^{T}$$

- Any doubts ???

Thank you!