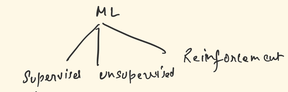
Introduction

Sunday, April 14, 2024 11:38 PM



↓

$D \in [x, y)$

* x : → features
  y : → labels
  → How to identify?

$y \in \mathbb{R}$ { 50, 60, 70, 30 } → Regression

$y \in [0, 1, 2, \ldots 10]$ → Classification

*

Text (String)	Label
M_1 → Messages	Spam
M_2 → "	not Spam
M_3 → "	Spam
M_4 → "	Spam

String

[Spam, non-Spam]

0 1

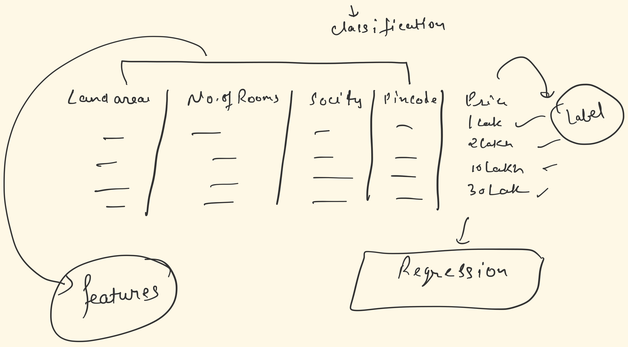
↓

Discrete value

↓

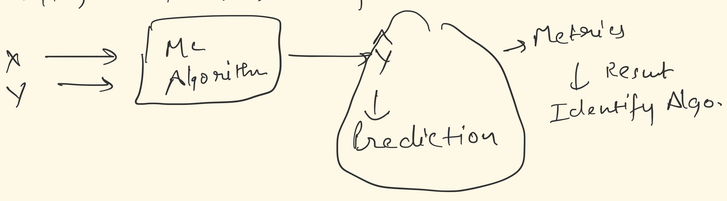
classification

*



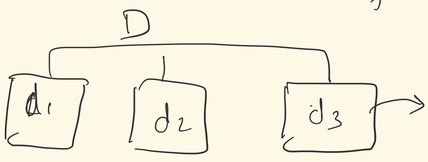
M.L training

* Everything should be numeric
* Data is always non-Linear in Real World
* find feature → $(X) \in \mathbb{R}^D$
  ↳ [(10 × 5)]
* find Labels $y \in \mathbb{R} / [0,1]$
* Problem ∈ [Regression, classification]
* Apply multiple Algo and find the best



* Text-preprocessing

1) TF-IDF (term frequency, Inverse document frequency)



Term frequency

$tf(t, d_2) \Rightarrow \frac{5}{25} = \left(\frac{1}{5}\right) = .2$

↓

term

↓

word.

$\Rightarrow \frac{20}{25}$

$$\frac{\text{total time word occure in document d}}{\text{total number of word in document d.}}$$

is, the, a

Inverse document frequency

$$IDf(t, D) \Rightarrow \left(\frac{N}{\text{number of time term occure in docs}}\right) \log$$

$$\Rightarrow \frac{3}{3} \rightarrow \log\left(\frac{3}{3}\right) \rightarrow 0$$