Data Sciene and Machine Learnig

Brahim Zirari[1]

Lectures

[1] PhD in economic analysis and forecasting. Department of economics. University of Medea. Email: ziraribr@gmail.com.


data
science
research
scientists
support
environment
faculty
students
activities
fellows
sciences
program
institute
bridge
space
education
uw
project
escience
incubator
activity
graduate
director
evaluation
metrics
methodology
reproducibility
engineering
budget
career
university
one
projects
studio
tools
committee
social
across
scientific
areas
statistics
physical
center
working
also
researchers
section
collaborations
effort
group
administrative
community
appendix
professor
theme
computer
new
work
funded
programs
initiative
development
number
software
paths
participate
training
time
year
impact
campus
established
establish
participants
people
staff
collaboration
fellow
courses
postdocs
flagship
ethnography
steering
washington
domain
alternative
open
described
breakthrough
see
use
metrics
funding
annual
anticipate
methodologies
long-term
including
environments
libraries
events
fields
success
online
grant
successful
goal
team
range
many
executive
leaders
methods
proposal
culture
universities
positions
three
techniques
current
among
office
focused
area
additional
practices
computing
seed
common
visiting
aligned
include
per
nsf
best
igert
assessed
well
permanent
years
funds
made
groups
process
external
individuals
roles
annually

1. What is Data Science?
2. What is a dataset?
3. Types of Data?
4. Machine Learning
5. Supervised Learning (Regression)

What is Data Science?

نبذة عن تخصص علم البيانات

عرّفت مجلة Harvard Business علم البيانات أنّه من أكثر العلوم إثارة وشيوعًا في القرن الحادي والعشرين، وأشار العديد من الخبراء في علم البيانات أنّ التشويق الذي يتمتّع به علم البيانات قد أضاف لمسة مثيرة إلى بعض العلوم الأخرى كالإحصاء مثلًا!

هل تعلم أن اللقب "نفط القرن الحادي والعشرين" يُطلق على تخصص "علم البيانات" وذلك لإبراز أهميته وقيمته العلمية في حياتنا؟

يُعتبَر علم البيانات أنّه علم متعدّد المجالات، كما أنّه العلم الذي يستخدم الأساليب العلمية، والعمليات، والخوارزميات، والأنظمة بغرض استخراج المعرفة، والأفكار من البيانات سواءً كانت هذه البيانات منظّمة ام لا، تمامًا مثل التنقيب عن البيانات وما يُطلق عليه "Data Mining".

WHAT IS DATA SCIENCE?

- "**Data science**, also known as **data-driven science**, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."
- "Data science intends to analyze and understand actual phenomena with 'data'. In other words, the aim of data science is to reveal the features or the hidden structure of complicated natural, human, and social phenomena with data from a different point of view from the established or traditional theory and method."

WIKIPEDIA
The Free Encyclopedia

What is a Dataset?

A dataset is a collection of data, typically organized in a structured format. Datasets can include a wide range of information, such as numerical values, text, images, or audio recordings. They are mostly used in fields like machine learning, business, and government to gain insights, make informed decisions, or train algorithms.

مجموعة من البيانات المنظمة في صيغ معينة و هي إما تكون هيكلية أو غير هيكلية.
تعتبر البيانات حجر الأساس في علم البيانات و تعلم الآلة.

Task:

1. With your team, use AI tools to figure out the difference between a dataset and a database.
2. What defines a good prompt?

Types of Datasets?

Numerical or quantitative Dataset

Ex:
Marks
Temperature

Categorical or qualitative Dataset

Ex:
True/False
Eye colour

Multivariate Dataset

Ordered Dataset

Ex:
customer reviews

Time series Dataset

Ex:
GDP growth

Image Dataset

Ex:
types of diseases

Types of Data ?

✓ Structured and Unstructured data

(بيانات هيكلية و بيانات غير هيكلية)

✓ Labeled and **Unlabeled data**

بيانات تحتوي على مدخلات و مخرجات و **بيانات تحتوي على مدخلات فقط**

✓ Training data and Testing data

بيانات التدريب و تتستعمل لبناء الخوارزمية، أما بيانات التقييم فتكون بغية اختبار مدى نجاح الخوارزمية

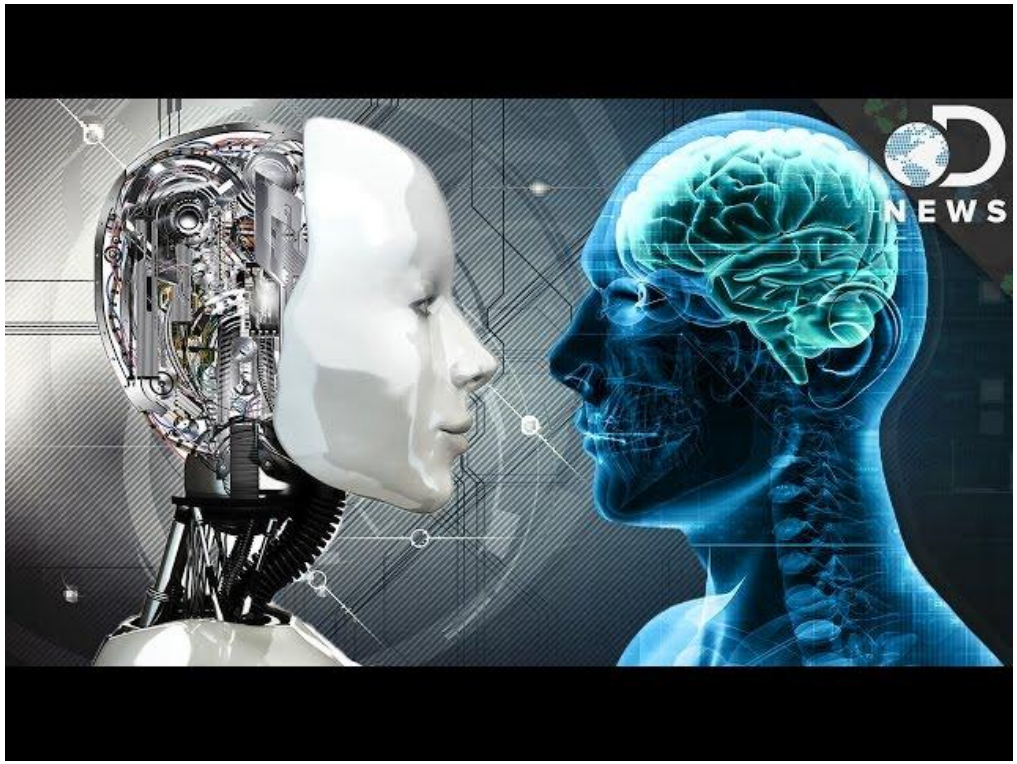
Machine Learning

Artificial Intelligence

قدرة الآلة على اتخاذ القرار المناسب

The ability of taking the right decision

How Human Thinks?

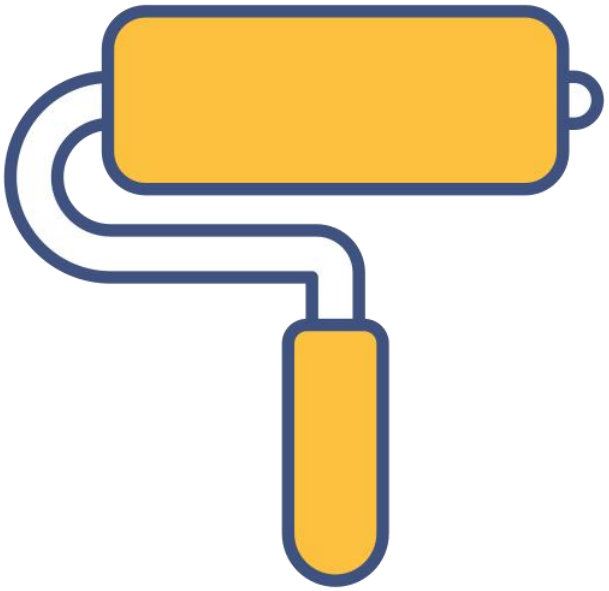


How Human Thinks?

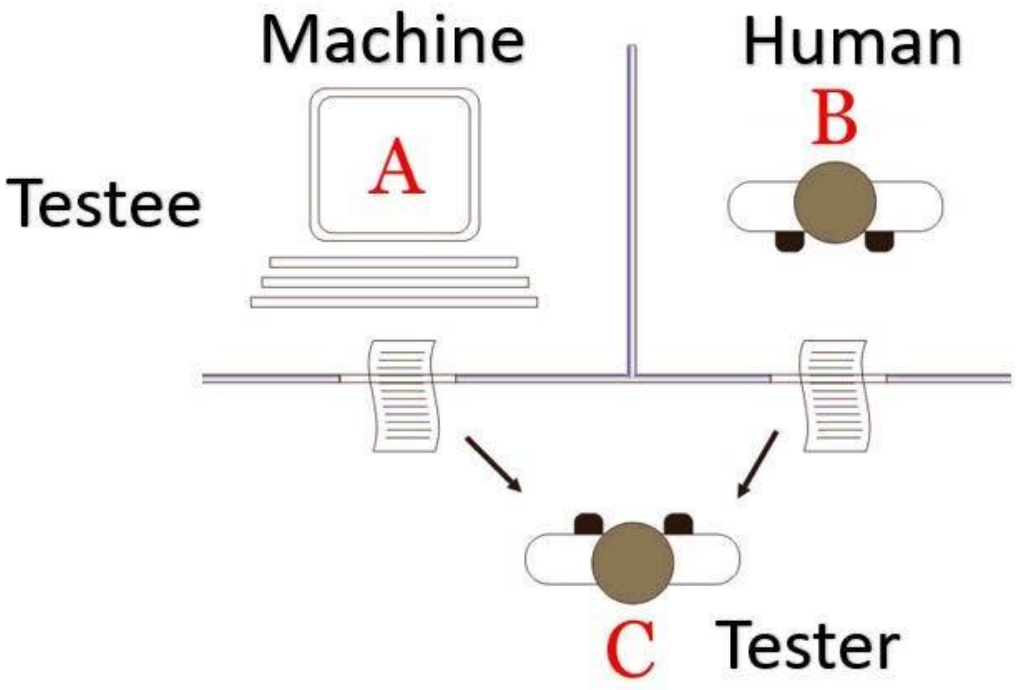
Model 1: Punishement and Reward

How Human Thinks?

Model 1: Supervision



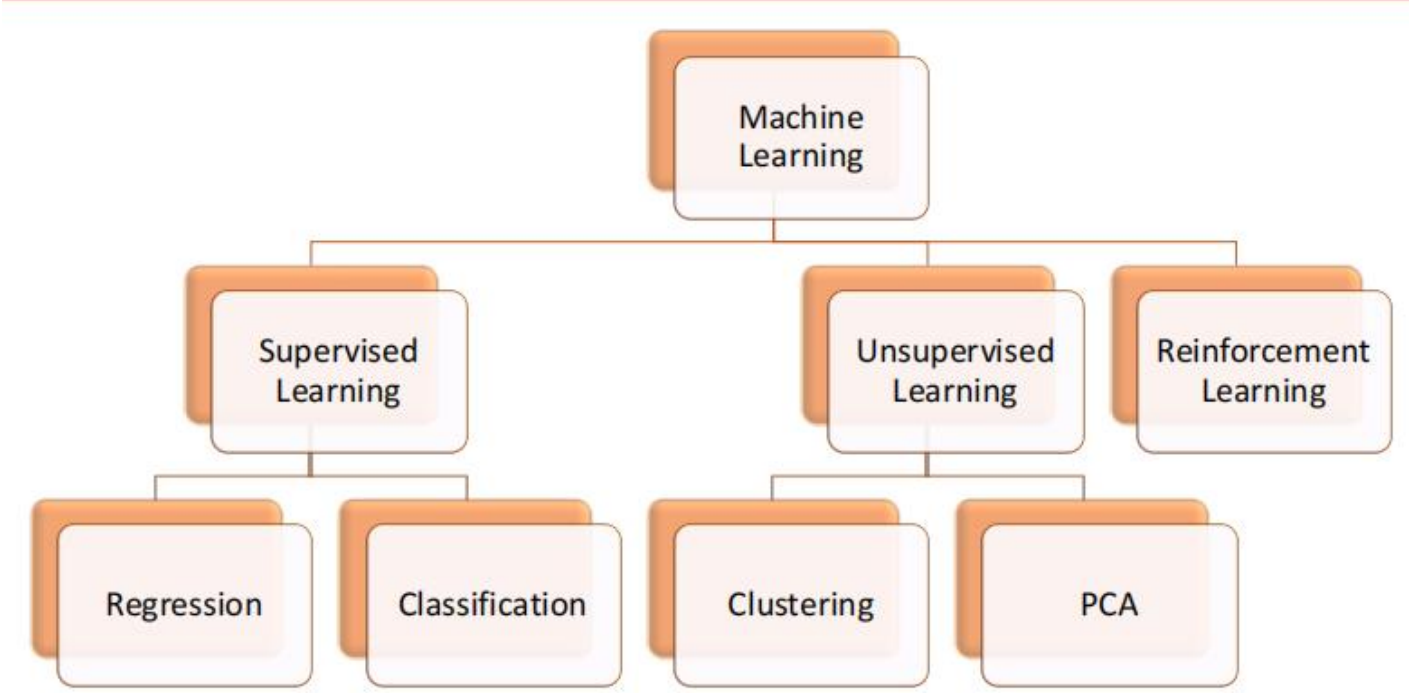
Turing Test



Machine learning is a subfield of computer science that deals with the design and development of algorithms that allow computers to learn from data without being explicitly programmed

القدرة على اتخاذ القرار المناسب بعد التدرب على بيانات

Machine Learning Models



Supervised Learning

Regression

Labeled data

Quantitative
OUTPUT

Classification

Labeled data

Categorical
OUTPUT

ML Vocabulary

Input	X
Output	Y
Rows	m
Features	n
h(x)	Prediction function
Cost J	Cost function
Theta	Weights

Regression with one feature (variable)

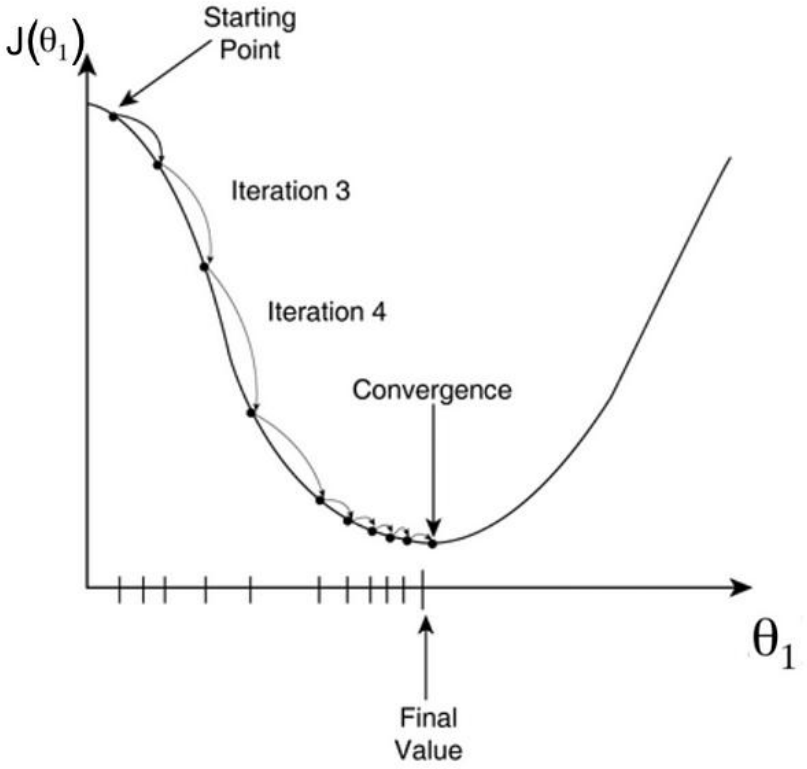
Hypothesis: $h_\theta(x) = \theta_0 + \theta_1 x$

Parameters: θ_0, θ_1

Cost Function: $J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$

Goal: $\underset{\theta_0, \theta_1}{\text{minimize}}\ J(\theta_0, \theta_1)$

Gradient Descent



Cost Function – "One Half Mean Squared Error":

$$J(\theta_0, \theta_1) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

Objective:

$$\min_{\theta_0, \theta_1} J(\theta_0, \theta_1)$$

Derivatives:

$$\frac{\partial}{\partial\theta_0}J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\frac{\partial}{\partial\theta_1}J(\theta_0, \theta_1) = \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)\cdot x^{(i)}$$

Example :

Theta 0= 5, Theta 1= 2, Alpha=0.01

X	Y	h(x)	h(x)-y	$(h(x)-y)^2$
1	7	7	0	0
2	8	9	1	1
2	7	9	2	4
3	9	11	2	4
4	11	13	2	4
5	10	15	5	25
5	12	15	3	9

Cost J= 3.3, **Convergence**!!

Example :

Cost J= 3.3, **Convergence**!!

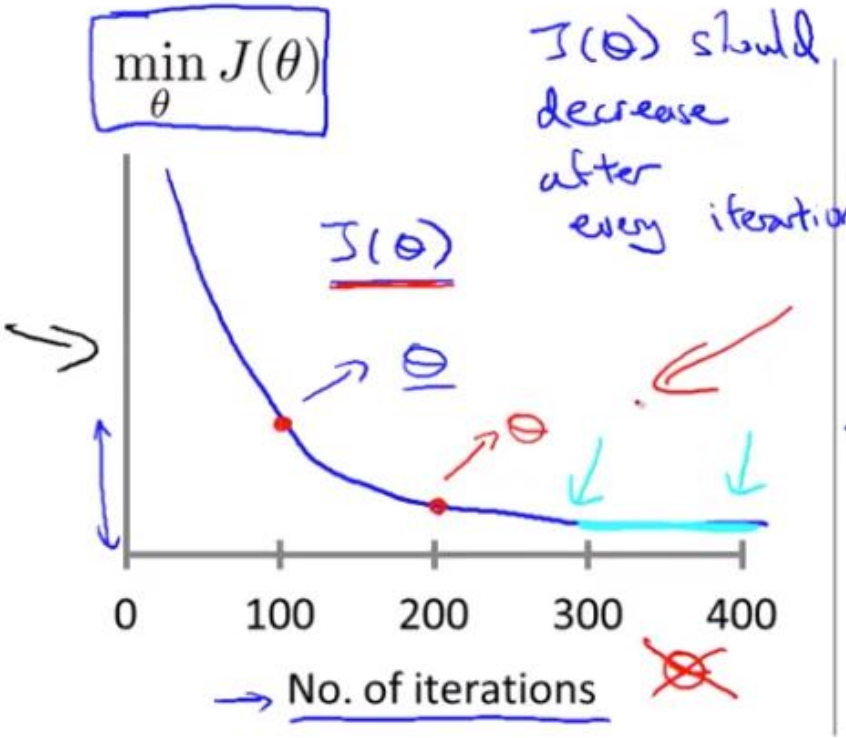
$$\text{repeat until convergence } \{$$
$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)$$
$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) \cdot x^{(i)}$$
$$\}$$

ϕ_0 = 4.993, ϕ_1 = 1.48....etc **How many times?**

Number of Iteration

عدد المحاولات

Making sure gradient descent is working correctly.



→ Example automatic convergence test:

→ Declare convergence if $J(\theta)$ decreases by less than 10^{-3} in one iteration.

ε

Example:

Theta0= 5, **Theta1**= 2, **Theta2**= 3, **Theta3**= 6, **Alpha**=0.01

X1	X2	X3	Y
5	20	6	114
5	35	6	120
6	38	8	123
7	40	8	121
7	46	10	135

أحسب قيمة الخطأ

بالاعتماد على خوارزمية الانحدار التدريجي أحسب قيمة ϕ_1 باستخدام محاولة واحدة

GD algorithm

Repeat until convergence

{

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

j= 0,...,n

}

Normal Equation

$$\theta = (X^T X)^{-1} . (X^T y)$$

Singular matrix

Normal equation vs Gradient Descent

1. do no require iterative approach;
2. do not require a learning rate;
3. do not require scaling.

Example with Python

1. Regression with one variabe;
2. Multivariable Regression;
3. Normal Function;