

Course Introduction

Dr. Mohamed Elshenawy



1

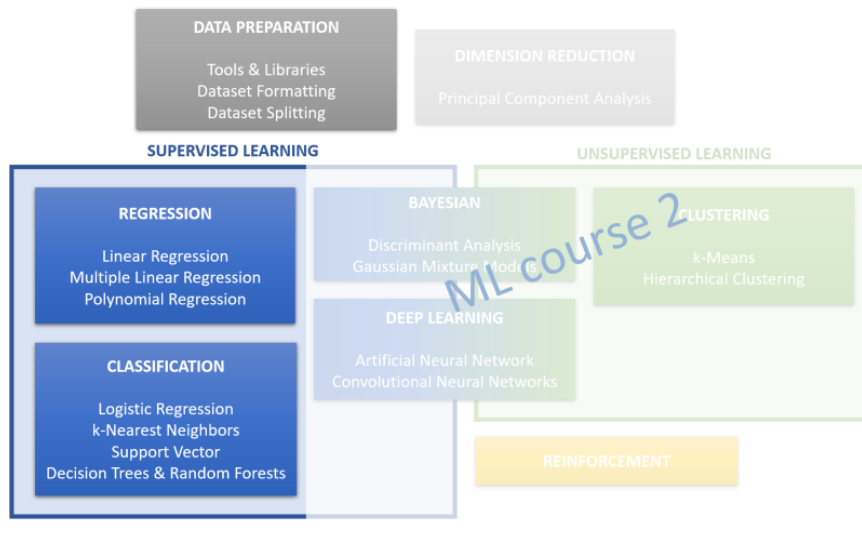
In this session ...

- Course Overview
- What is Learning?
- Why do we need machine learning?
- Types of learning
- Practice
 - Learn about the tools and libraries
 - Start playing with the data
 - Explore the dataset
 - Understand different data types
 - Pre- Processing



2

Scope



3

Course Overview

- 5 days and 6 Sessions
- Day 1
 - **Session one:** Introduction to Machine Learning, tools and data preparation, and introduction to linear regression
 - **Session two:** Multiple and Polynomial Regression, [assignment 1](#)
- Day 2
 - **Session one:** Assessing the performance of a regression model, Project 1
 - **Session two:** Classification, K-NN, and Logistic Regression, [assignment 2](#).



4

Course Overview

- 5 days and 6 Sessions
- Day 3
 - **Session one:** Assessing the performance of a classification model, and Project 2.
 - **Session two:** SVM and SVR, [assignment 3](#)
- Day 4
 - **Session one:** Decision Trees and Random Forest
 - **Session two:** Ensemble Learning, [assignment 4](#)



5

Course Overview

- Day 5
 - **Session one:** Model selection and hyperparameter tuning, introduce the challenges
 - **Session two:** Solve the challenges



6

Let's break the ice

- About me.
- How about you?
- What do you know about course topics?
- What do you expect from me?



7

Why do we need to take this course?

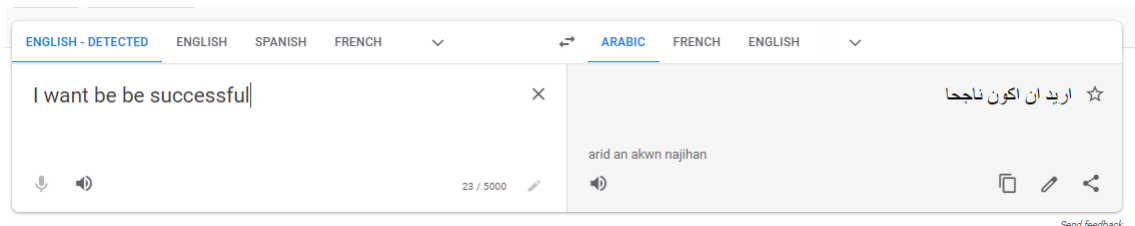
- What type of problem can I solve after taking this course?
- It is a first step to create a **smart** system.
- What do you mean by **smart**?



8

Smart System?

A system that can detect a language of a sentence and translate it to another language.



9

Smart System?

A system that can help choose the products to buy

Frequently bought together



+



+



Total price: \$170.60

Add all three to Cart

Some of these items ship sooner than the others. Show details






10

Smart System?

A system that can help you make better decisions

Restaurant Sales Testing > Sales Data.csv > dataset

rows: 600 columns: 7

view as:   

Menu Id	Date	Sales	Weather	Holiday	Festival	Promotions
G1	2017-01-01T00:00:00	104	0	0	1	0
G1	2017-02-01T00:00:00	112	0	0	0	0
G1	2017-03-01T00:00:00	115	0	0	0	0
G1	2017-04-01T00:00:00	252	1	0	0	1
G1	2017-05-01T00:00:00	132	1	0	0	1
G1	2017-06-01T00:00:00	110	1	0	0	1
G1	2017-07-01T00:00:00	173	1	0	0	1
G1	2017-08-01T00:00:00	232	1	0	0	0
G1	2017-09-01T00:00:00	172	1	0	0	0
G1	2017-10-01T00:00:00	215	0	0	1	0
G1	2017-11-01T00:00:00	304	0	0	1	0
G1	2017-12-01T00:00:00	68	0	1	1	0
G1	2017-01-01T00:00:00	104	0	0	1	0

slideshare.net

C1	C2	C3
Year	Advertising Cost (\$)	Yearly Sales Volume (Units)
2000	22450	120231
2001	23890	124432
2002	24540	125568
2003	24990	125897
2004	25230	126302
2005	25975	126906
2006	26450	127212
2007	26989	127342
2008	27123	127879
2009	27350	128213
2010	28110	128567
2011	28879	128909

<http://www.exodus-digital-marketing.co.uk/>



11

Smart System

- Helps you:
 - Filter unwanted emails
 - Detect harassing tweets/ messages.
 - Find fraudulent transactions.
 - And much more...



sendpulse.com

<http://www.efrongroup.com/>



12

Can we just write a typical computer program to solve these tasks?

- You will face two problems:
 - **Complexity**
 - It will be very difficult to specify a set of rules that solve problems such as translation, spam filters, and product recommendation.
 - **Adaptivity**
 - Many tasks change over time or from one situation to another (think of spam filters).



13

How can we implement all these nice features

- Using one tool

Machine Learning



14

A project I was part of (deep learning)



Youssef Y, Elshenawy M. Automatic Vehicle Counting and Tracking in Aerial Video Feeds using Cascade Region-based Convolutional Neural Networks and Feature Pyramid Networks. *Transportation Research Record*. March 2021. doi:[10.1177/03611981211997833](https://doi.org/10.1177/03611981211997833)



15

What is Machine Learning?






- “A computer program is said to learn from **experience E** with respect to some class of **tasks T** and **performance measure P** , if its performance at tasks in T , as measured by P , improves with experience E .” Mitchell (1997)
- **The Experience, E :** Most machine learning algorithms simply experience a dataset.
- **Tasks:** classification, regression, transcription, machine translation, anomaly detection, etc.
- **Performance:** specific to the *task T* (e.g. accuracy, mean square error, recall,)



16

Types of Learning – Supervised

- The machine learns by example
- You have a training data set in which the **correct output is known**

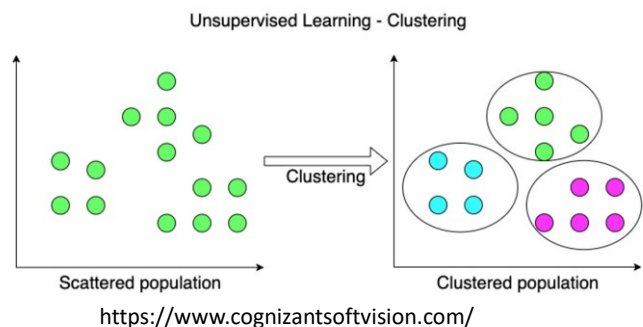
Input	Output (label)
	Cat
	Cat
	Dog
	Dog
	??



17

Types of Learning – Unsupervised

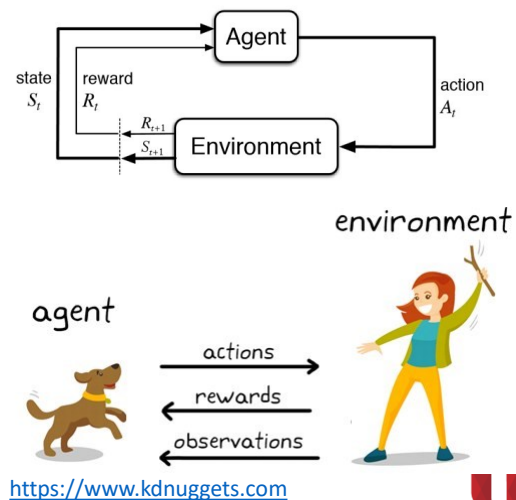
- Explore the dataset
- No labels/ desired outcome
- Discover regularities and hidden structures in data.
- Correct output is not known.
- Example applications
 - **Clustering**: organize instances into subgroups or clusters without any upfront knowledge.
 - **Dimensionality reduction**: extract the most relevant information/features.



18

Types of Learning – Reinforcement Learning

- Learn by practice
- An intermediate setting between supervised and unsupervised learning
- The goal is to learn through actions. A reward signal indicates if the action is good or bad.
- By interacting with the environment, the agent learns how to maximize the reward function.



19

About Learning: lessons from animal behavior

- Bait Shyness (Rats learn to avoid poisonous baits)
 - Rats will eat very small amounts of the food items with novel look or smell
 - When the food produces an ill effect, the novel food will be associated with the illness and the rats will not eat it. (learning by memorization)
- In another set of experiments carried out by Garcia, J. & Koelling, R. (1996), they found that using unpleasant stimulus such as electrical shock, do not cause the rats to avoid the food.

from the book "Understanding Machine Learning: From Theory to Algorithms." By Shai Shalev-Shwartz and Shai Ben-David



20

Pigeons

- In an experiment conducted by the psychologist Burrhus Frederic Skinner:
 - A number of hungry pigeons were placed in a cage.
 - each pigeon was engaged in some activity (pecking, turning the head, etc.) to find food.
 - Food is delivered to the pigeons at regular intervals (it does not depend on their behavior)
 - He found that each pigeon tends to spend more time doing the same action that he was doing when the food arrives which increases the chance that the pigeon will be doing the same activity when the new food arrives.
- Inductive reasoning might lead us to false conclusions.



21

Why rats' learning more successful than that of the pigeons?

- While the pigeons, in the experiment, are willing to adopt any explanation for the food arrival, rats know that food cannot cause electrical shock.
- The rats seem to have a prior knowledge that it is unlikely that food consumption and electrical shocks have a causal relationship (*inductive bias*).
- Assessing our models is not always straight forward
- Correlation does not imply causation



22

References

Books

- An Introduction to Statistical Learning. James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani, 2013, ISBN: 978-1-461-47137-0.
- Pattern Recognition and Machine Learning for Christopher Bishop.

Courses

- Andrew Ng's Machine Learning
 - <https://www.coursera.org/learn/machine-learning#about>
- ML course of Prof. Ghassany
 - <https://www.mghassany.com/MLcourse/>

Python

- Crash Course on Python.
 - <https://www.coursera.org/learn/python-crash-course>



23

Let's start



24

Building Machine Learning Models

- To build a machine learning model, you need....

Data



25

Where can I get data to build my model?

- For practice, you have many sources:
 - Kaggle: <https://www.kaggle.com/datasets>
 - UCI Machine Learning Repository: <https://archive.ics.uci.edu/ml/datasets.php>
 - Registry of Open Data on AWS: <https://registry.opendata.aws/>
 - Google <https://datasetsearch.research.google.com/>
 - More <https://github.com/awesomedata/awesome-public-datasets#machinelearning>



26

How about real-world projects

- For an innovative project:
you need to be innovative 😊 --
- Collect data yourself, simulations, use existing data in a different way, ..
find a way
- In many cases, you may deal with legacy systems and extract data from
them



27

- Now I have the data, can I build the model?
- Not yet, you have to prepare the data



28

Prepare your data

- Unfortunately, it is uncommon that you will use the data as it is. You need to:
 - **Clean your Data:** identify and handle errors in your data.
 - **Data Transformation:** change the scale of some/all variables. Why?? We will discuss it later.
 - **Feature Selection:** select these features that are most relevant to your task.
 - **Feature Engineering:** combine features, derive new variables, dimensionality reduction, etc.



29

In the practice session

- Libraries are will use frequently in this course
 - Pandas: Data analysis and manipulation tool
 - Numpy: simplify working with multi-dimensional arrays and matrices.
 - Matplotlib: plotting library
 - Scikit-learn: create machine learning models
 - Seaborn: advanced visualization functions
- Working with datasets (no modelling yet 😊)
 - Import
 - Slice
 - Explore
 - Visualize
 - Some common pre-processing



30

Source of today's practice sessions

- <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/overview/tutorials>

