Advanced Topics in Information systems

SE204

Lecture 7

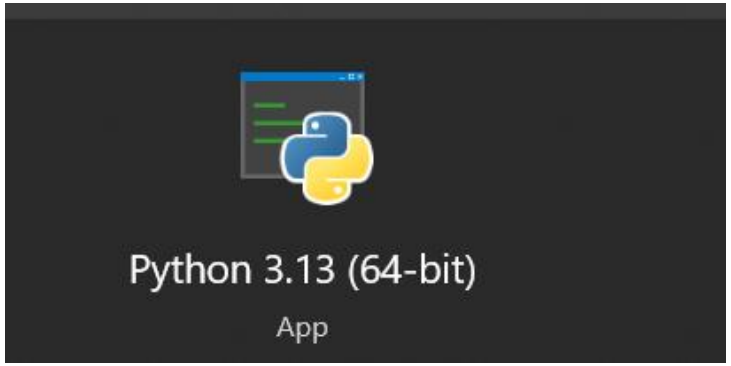
Dr. Nelly Amer

Agenda

- **Python for frequent pattern mining cont.**
- **What is machine learning.**
- **Machine learning and interpretability.**
- **Types of machine learning.**
- **Machine learning algorithms.**

Python cont.

After finish setup, In start menu you have the following:



- This is python shell
- you run one line at a time, you can not save command
- You can access it according to its icon in start menu or by CMD (enter CMD and write python)

- python idle =Python's Integrated Development and Learning Environment)
- IDLE provides a simple text editor where you can write and save full Python programs (.py files).

This tool allows you to read Python's official documentation, including information about built-in modules, functions, and classes.

Libraries install

To install any library

From cmd you can install libraries, not from idle, not from shell(python 3.13(64-bite))

Type:
pip install library name

Examples:

pip install pandas → Creating a DataFrame (Like an Excel Table)
pip install mlxtend → supports both Apriori & FP-Growth

How to Write and Run a Script in IDLE

1- From the Start Menu Open **IDLE (Python 3.13 64-bit)**.

2- Click **File → New File**.

3- Type your Python code, as:

```
print("Hello World")
x=5+3
print("x=",x)
```

4- Save the file from **File → save**(Ctrl + S) with a **.py extension** (e.g., test.py).

5- Run the script by clicking **Run → Run Module** (F5).

6- The output will appear in the **Python Shell (interactive window)**.

How to Write and Run a Script in IDLE cont.

Practical examples:

Open IDLE (Python 3.13 64-bit) and writ examples includes:

- Import libraries
- Input data
- Convert it to Boolean
- Use Apriori and fpgrowth functions, knowing the meaning of their parts
- Determine strongest association rules, according to minsupport and minconfidence
- How to handle NaN, and its reasons.
- Print rules
- Python case-sensitive.

What is machine learning.

Machine learning (ML)

Machine learning (ML) is a field of computer science that studies algorithms and techniques for automating solutions to complex problems that are hard to program using conventional programing methods.

Machine learning focuses on creating predictive models that learn from data and generalize unseen data

I.e., The larger the dataset, the more accurate they become

The conventional programming

The conventional programming

consists of two distinct steps:

The first step: is to create a detailed design for the program, i.e., a fixed set of steps or rules for solving the problem.

The Second step: is to implement the detailed design as a program in a computer language.

Machine learning vs conventional programming

The conventional programming

Design a program that sum 2 integers

input X , Y

Z= x+ y

Output z

You give the computer **Rules** , **Inputs** (X, Y)

The computer applies the rules and gives **output**

i.e., you **tell the computer what to do**.

Machine learning (ML)

Given the input data, and output data, design a model to predict the output for new input data.

Input data=[(1,3), (4,7), (8,12),……(7,20)]

output data=[(4), (11), (20),……(27)]

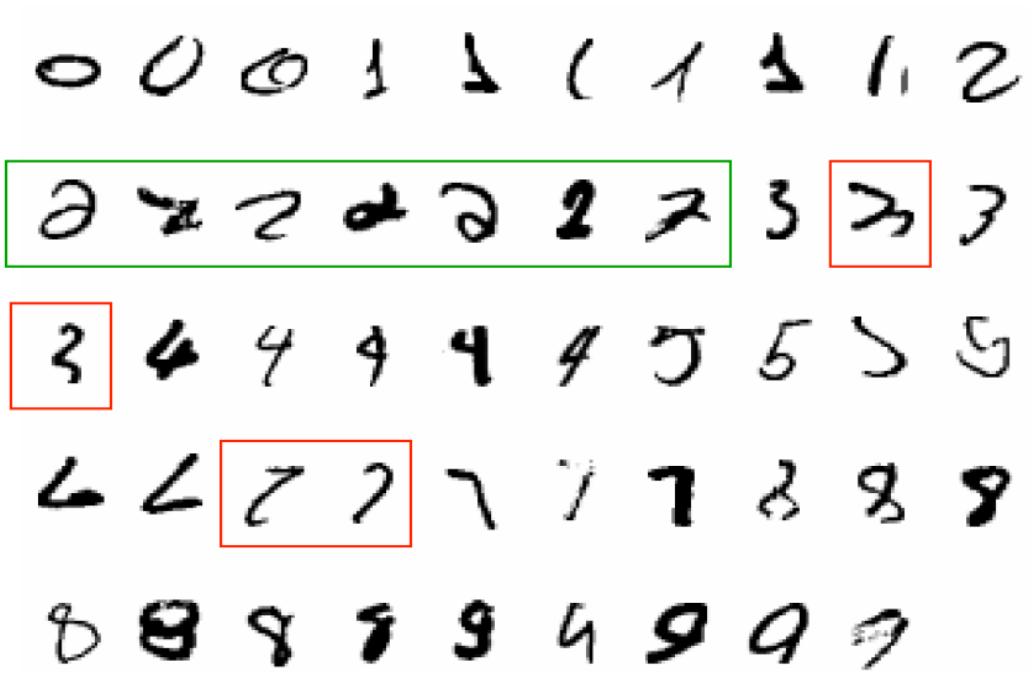
Machine learning algorithm find the relation between input and output

You give the computer **Input–Output pairs** (examples)

The computer finds the **rules** by itself

i.e., you **give it examples**, and the computer **figures out what to do**.

Machine learning cont.



Looking at the picture, decide which is a cat and which is a dog

given a picture, determine which symbol makes 2

This could be done :

By machine learning ✓

By conventional programming ✗

Machine learning cont.

ML research made slow and steady progress on solving complex problems until the mid-2000s, after which the progress in the field accelerated drastically. The reasons for this dramatic progress include:

- Availability of large amount of data due to the Internet, such as large datasets of images
- Availability of large amounts of compute power, supported by **large memory and storage space**
- Improved algorithms that are optimized for large datasets.

advancements in ML include the following:

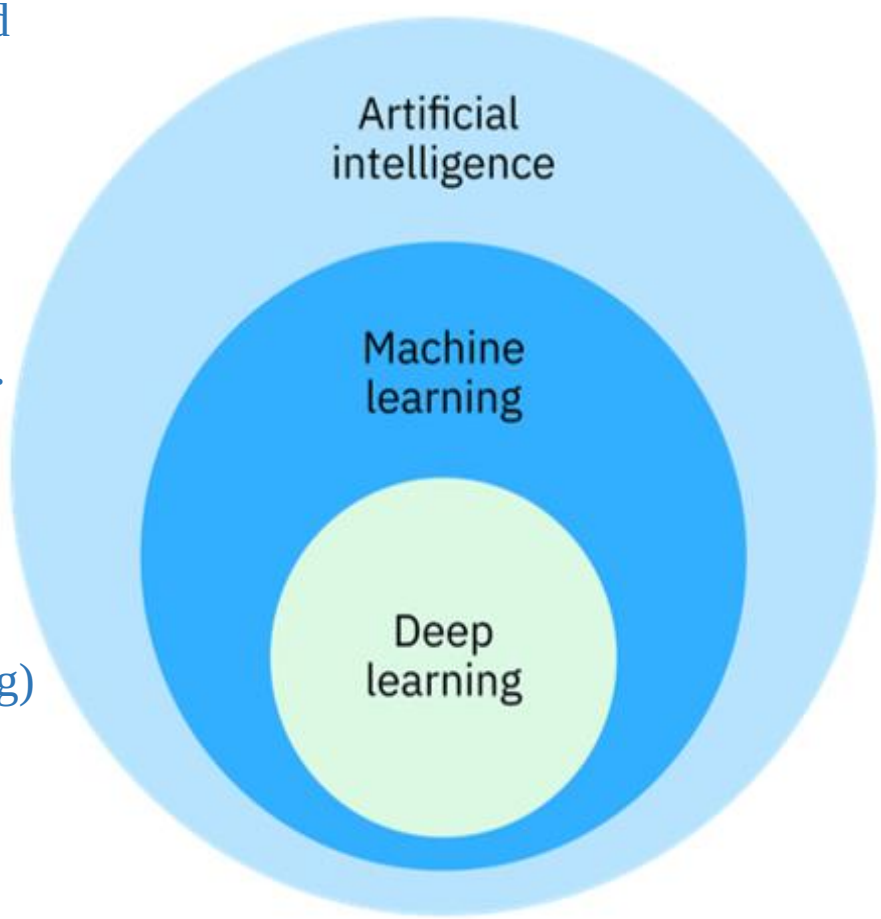
- ❑ Speech recognition – the ability to recognize speech and convert it to text.
- ❑ Language translation – understanding and forming language constructs without formal training in grammar, and a huge increase in the accuracy of translation compared to earlier methods.
- ❑ Driverless vehicles – the ability to navigate the vehicle without human intervention

Machine learning & Deep learning & Artificial intelligence (AI)

AI Goal: Create systems that can perform tasks that require human intelligence.
It enables computers to simulate human thinking, decision-making, learning, and problem-solving.

ML is a **subset of AI** that focuses on the idea that machines can **learn from data** and improve over time without being explicitly programmed for every task.

Goal: Build algorithms that can learn from data and make predictions or decisions.



Other techniques in AI are not based on ML:

- Expert Systems
- Traditional Computer Vision
- Game Playing
- Traditional NLP (natural language processing)

Deep learning is a subset of **machine learning (ML)** that uses **artificial neural networks with multiple layers** to model complex patterns in data.

Basic concepts

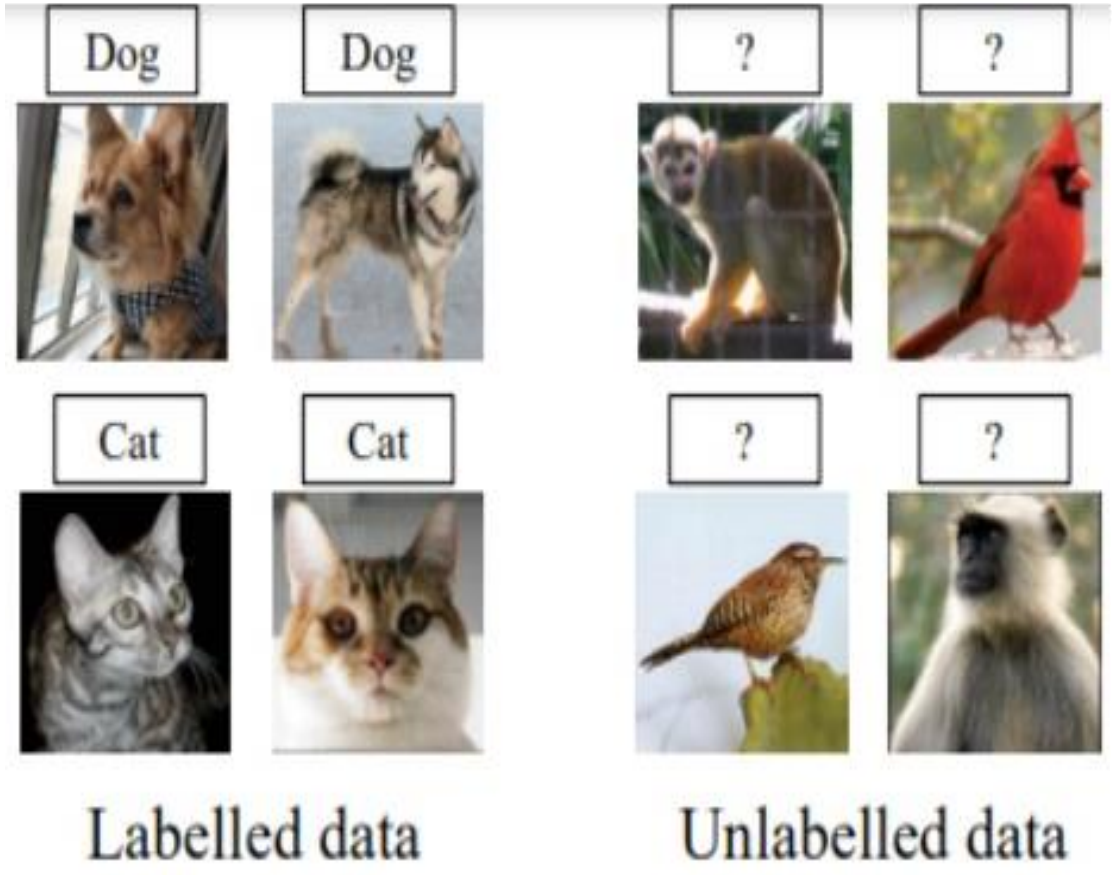
❑ Labeled data & unlabeled data

Labeled Data

Any data which has a characteristic, category, or attributes assigned to it can be referred to as labeled data. For example, a photo of a cat, a photo of dog.

Unlabeled data

Any data that does not have any labels specifying its characteristics, identity, classification, or properties can be considered unlabeled data. For example photos, videos, or text that do not have any category or classification assigned to it can be referred to as unlabeled data.



Basic concepts cont.

❑ Training set & testing set

Training data:

Training data is the data used for training the model.

Testing set:

Testing data is new data that the model **has never seen**. We use it to **evaluate how well** the model can generalize to new, real-world data.

❑ Overfitting & underfitting

Overfitting

Overfitting happens when a model learns the training data too well and doesn't generalize well to new, unseen data. It occurs when the model is very complex for the amount of training data given.

Underfitting

Underfitting happens when a model is too simple to learn the underlying pattern in the data. It performs poorly on both the training data and new, unseen data.

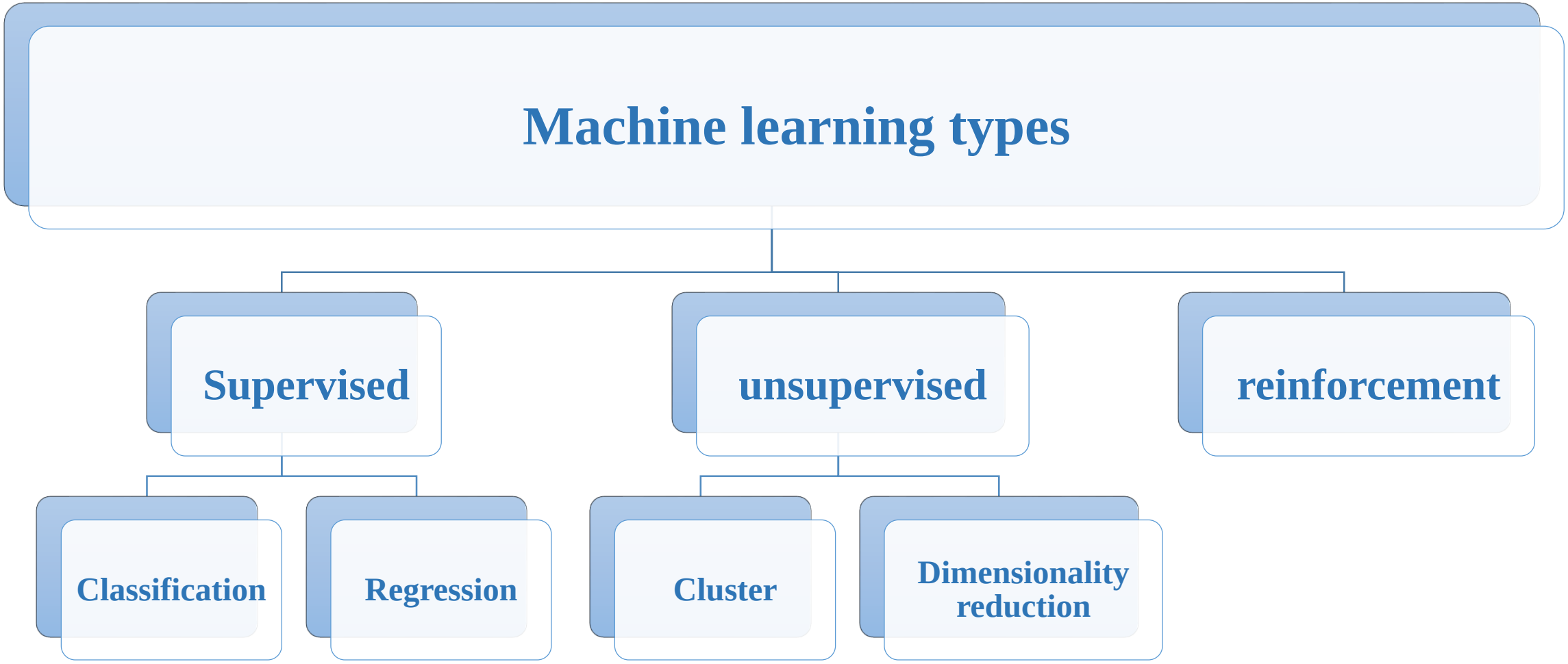
Machine learning types

Supervised learning → (classification, regression)

Unsupervised Learning → (cluster, Dimensionality reduction)

Reinforcement Learning

Machine learning types.



Supervised learning

Supervised machine learning is one of the most commonly used and successful types of machine learning.

It is applied when there is a need to predict a specific outcome based on input data, and we have examples of input and output pairs (labeled data).

In supervised learning, models are trained on labeled datasets, where each input is associated with a known output.

The goal is to learn a mapping function so the model can accurately predict outputs for new, never unseen inputs.

real-world applications of supervised learning :

Fraud detection
Spam filtering
Medical diagnosis
Email classification

Unsupervised learning.

Unsupervised machine learning is used when we do not have labeled data — that means we only have input data and no specific output.

The goal is to let the machine discover structures, or relationships in the data on its own.

Unsupervised learning is helpful when:
We don’t know the categories or groups in advance.
We want to group similar items.

The unsupervised learning algorithm looks at the input data and tries to group, cluster, or reduce dimensions based on similarities.

real-world applications of unsupervised learning :

Customer segmentation in marketing.
Organizing documents by topic
Reducing data for visualization

Reinforcement Learning

In the Reinforcement learning, an agent is interacting with an environment. The agent observes the state of the environment, and based on its observations, it can choose an action. Depending on the chosen action, it gets a reward. If the action was good, the agent will get a high reward and vice versa. **The goal of the agent is to find the best action for each state**.

How It Works

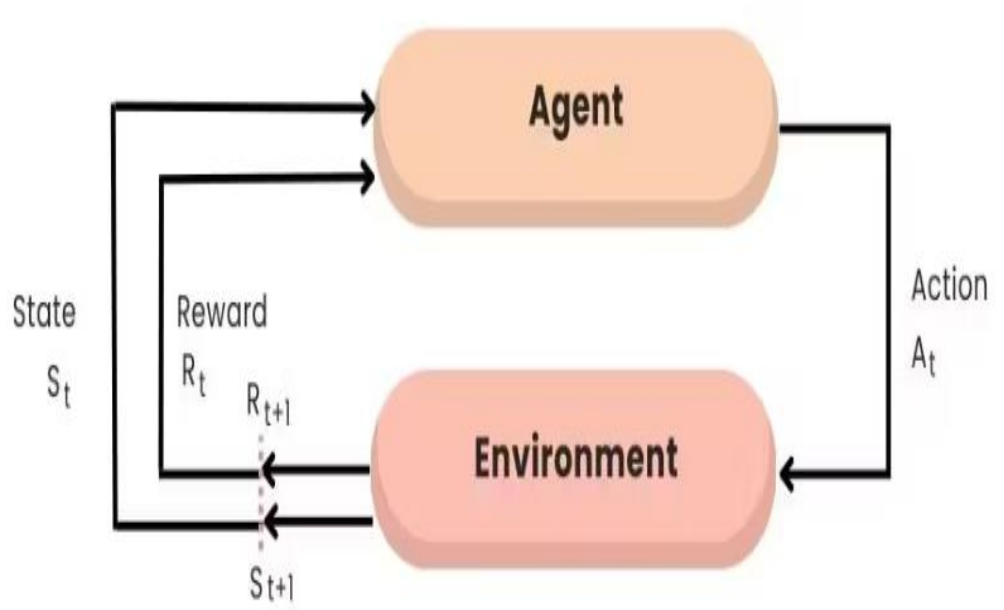
1- The agent observes the environment (where it is in a maze).

2- It takes an action (move left or right).

3- The environment gives a reward:
Positive: if it moves toward the goal
Negative: if it hits a wall

4- The agent learns: which actions lead to higher rewards.

5- Repeat: The agent keeps trying, improving its strategy over time.



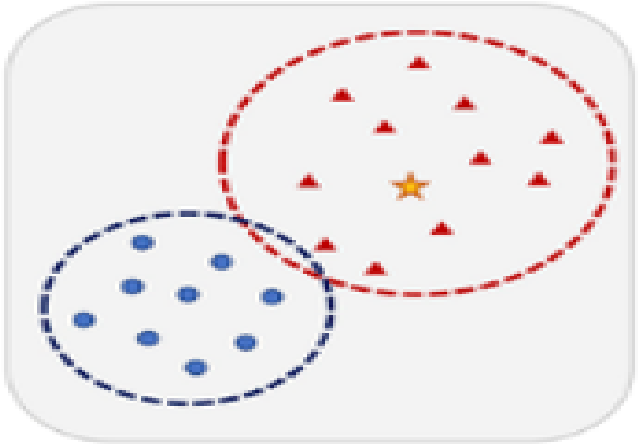
Example of reinforcement learning:

many robotics applications that learn how to walk

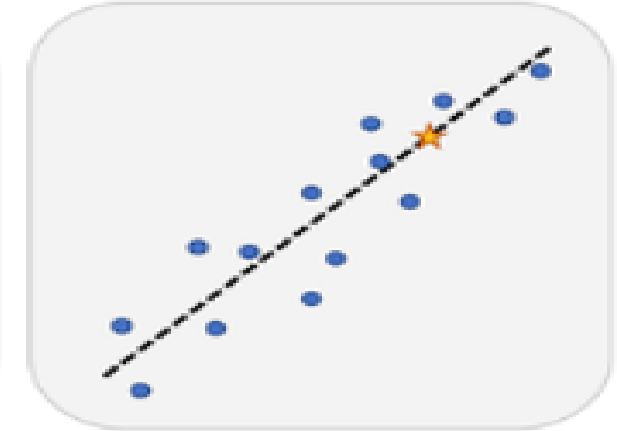
Supervised machine learning tasks

Supervised learning tasks

Classification



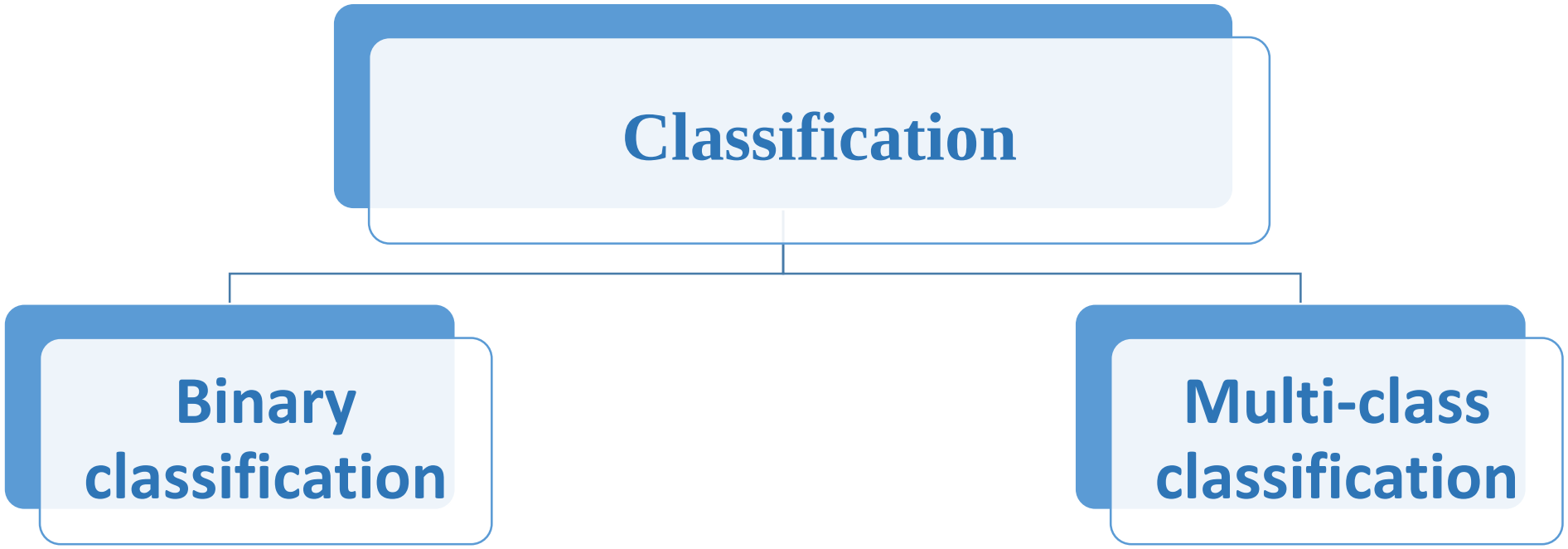
Regression



Classification machine learning.

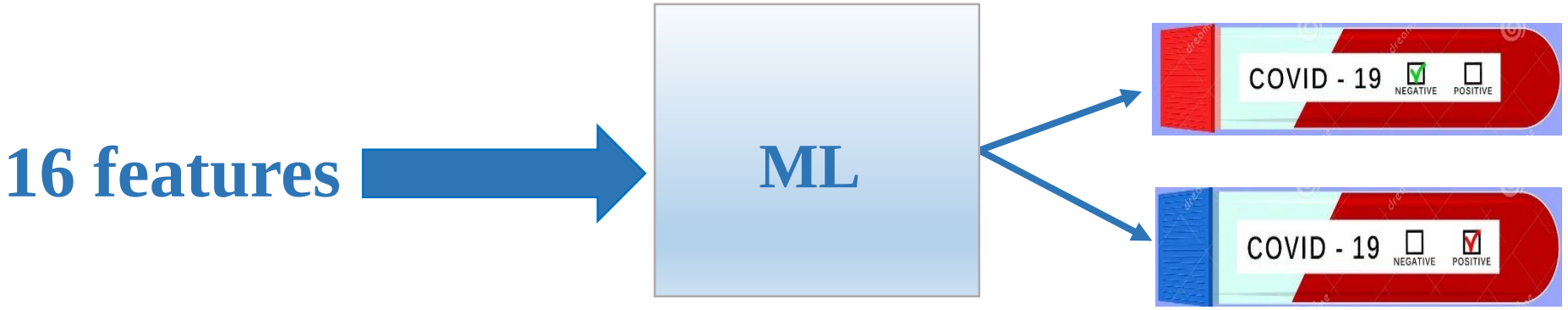
Classification is a supervised learning algorithm where a training set of labelled data is available. The model learned from training data to identify the category or class of the input feature or data is called classifier.

The classifier can be a binary classifier or a multi-class classifier.



Binary classification

A binary classifier identifies the input as belonging to one of the two output categories. For example, the mail received is a spam or not spam, Corona Virus Diagnosis classification.



Corona Virus Diagnosis classification.

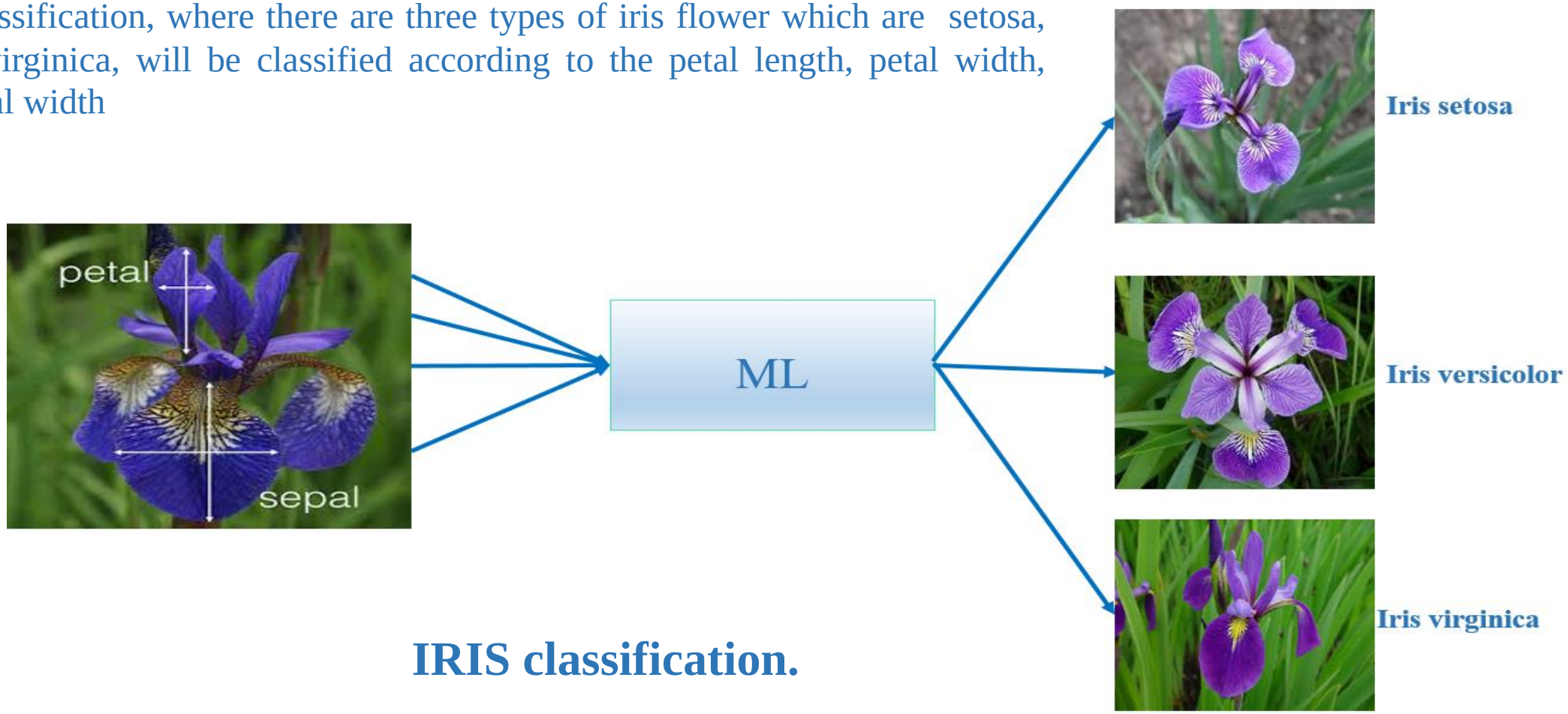
Multi-class classification

A multi-class classifier identifies the input vector as one of more than two categories.

For example:

1- the mail received is a promotional email that represents some kind of advertisement, personal email received from friends or associates, or a spam email.

2- Iris flower classification, where there are three types of iris flower which are setosa, versicolor, and virginica, will be classified according to the petal length, petal width, sepal length, sepal width



IRIS classification.

ROC Curve

The receiver operating characteristic (ROC) curve is a common tool used with **binary classifiers**.

The objective of the ROC curve is to evaluated the performance of a binary classification model.

How to Plot the ROC Curve:

The ROC curve is a graph with:

False Positive Rate (FPR) on the X-axis

True Positive Rate (TPR) (also known as Recall) on the Y-axis

Where:

FPR = FP / (FP + TN)= ratio of actual negative instances incorrectly classified as positive=

TPR = TP / (TP + FN)= ratio of actual positive instances correctly classified as positive

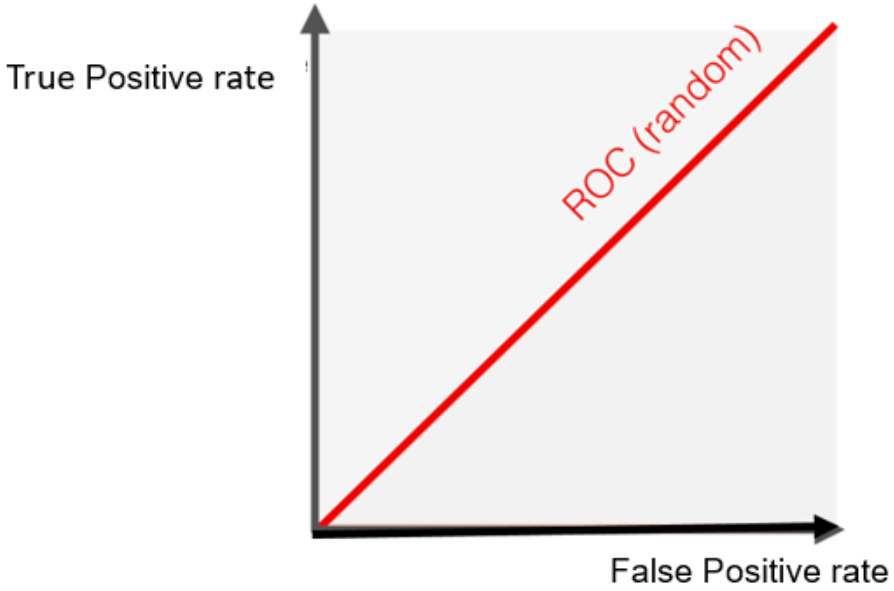
Types of ROC curve:

1- Diagonal ROC curve.

2- L-shaped ROC curve.

ROC Curve cont.

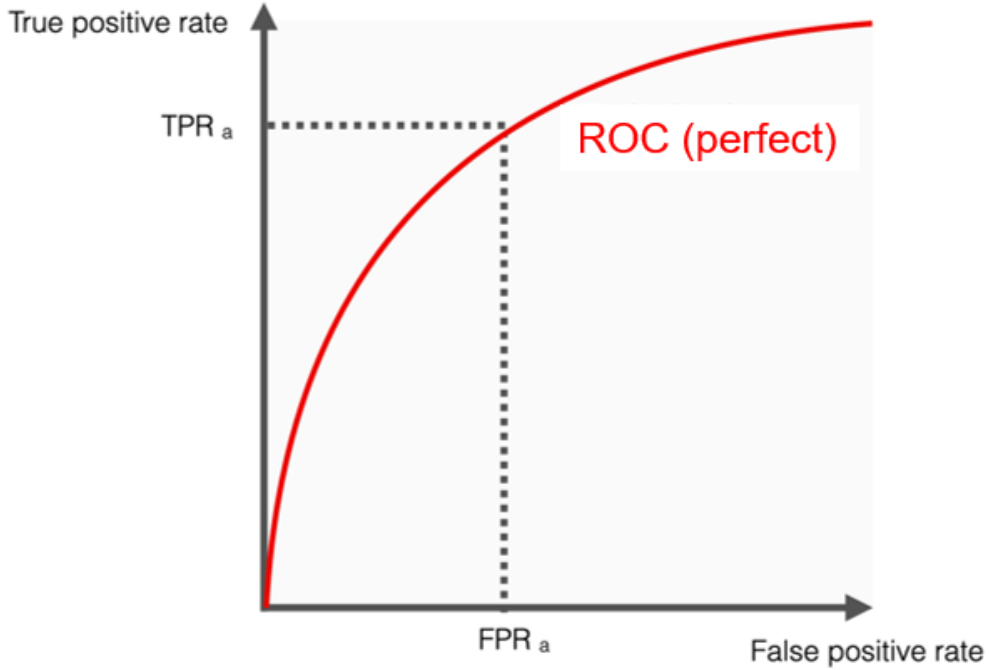
Diagonal ROC curve



Means the model performs like random guessing.
For every correct positive, there's a false one:
TP ≈ FP TN ≈ FN

It is bad model, since it cannot tell the difference between the two classes.

L-shaped ROC curve



Means the curve rises quickly toward the top-left corner.
Model makes many correct predictions:

- True Positives (TP) are high
- False Positives (FP) are low

It is great model — it separates the classes well.

Classification machine learning algorithms.

Classification algorithms

- **Logistic Regression**
- Decision Trees for Classification
- Artificial Neural Networks (ANNs) for Classification
- Support Vector Machines (SVM)
- Random Forest
- K-Nearest Neighbors (KNN)

We will focus on Logistic Regression

Logistic regression

Logistic Regression is a classification algorithm, which allows you to perform binary classification or multi-class classification.

For example:
Classify emails as spam or ordinary.
Classify iris flowers into Iris setosa, Iris virginica, or Iris versicolor.

In binary classification:
The sigmoid function is used to compute the estimated output $\hat{y}$:
$\hat{y} = \frac{1}{1+e^{-z}}$,
where $\hat{y}$ is the predicted output, and z is linear combination of features X (i.e., $Z=b_0+b_1x_1+b_2x_2+\ldots\ldots+b_n x_n$)

The performance of the classification model is evaluated during training using a cost function in the form of a loss function:
For binary classification, the loss function is:
function : $L(y, \hat{y}) = -[y * log(\hat{y}) + 1 - y(log(1 - \hat{y})]$.
where: y is the true label (either 0 or 1), $\hat{y}$ is the predicted probability (output of the sigmoid function).

In multi-class classification:
The softmax function is used to compute the probabilities for each class.

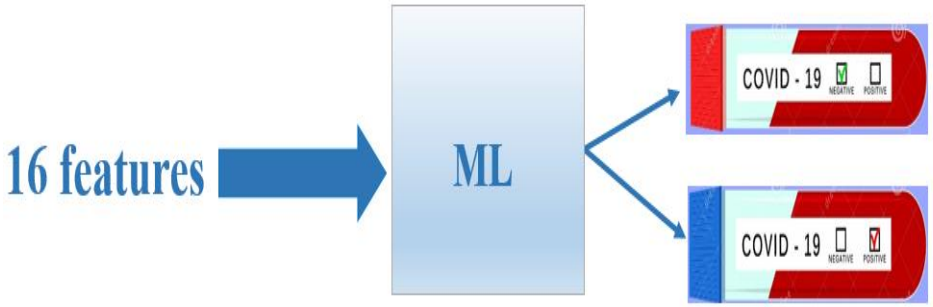
Logistic regression

We have input vector of X (features of blood analysis for 20 individuals)

and their Y vector output as +VE Corona virus or -VE Corona virus) :

$X=[(x_1,…..x_{16})_1, (x_1,…..x_{16})_2 ……,……..(x_1,…..x_{16})_{20}]$

Y=[-VE, +VE, ……...,+VE]



How the algorithm work:

1. Initialize the coefficients in the sigmoid function: $\hat{y} = \frac{1}{1+e^{-z}}$, where $\hat{y}$ is the predicted output, and z is linear combination of features X (i.e., $Z=b_0+b_1x_1+b_2x_2+……+b_n x_n$)
2. Enter The first instance.
3. Compute the predicted output $\hat{y} = \frac{1}{1+e^{-z}}$
4. Compute the error between the predicted output and the actual output using the cost function in the form of loss function : $\boldsymbol{L(y,\hat{y}) = -[y * \log(\hat{y}) + 1 - y(\log(1-\hat{y})]}$.
5. Update the coefficients of sigmoid function to minimize the error using **the gradient descent algorithm**.
6. Input the second instance and repeat the steps from 3 to 5 until end all instances as the first iteration,
7. Do more iterations until the model reaches convergence, hence reach the high accuracy.

Assignment #7

1- Given the following transactions, determine the strongest association rules with support 40% and confidence 60%., and print it, using python

tid	Set of items
1	$\{Bread, Butter, Milk\}$
2	$\{Eggs, Milk, Yogurt\}$
3	$\{Bread, Cheese, Eggs, Milk\}$
4	$\{Eggs, Milk, Yogurt\}$
5	$\{Cheese, Milk, Yogurt\}$

References.

[1] G. Rebala, A. Ravi, S.Churiwala “An Introduction to Machine Learning”, Springer, 2019

[2] A. Géron, “Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow”, O’Reilly, 2019