Machine Learning In Python

Subject: Unsupervised Learning

Lecturer : Reza Akbari Movahed

Hamedan University of Technology

Spring 2020

Unsupervised vs Supervised

Supervised Learning

- It contains training and testing steps.
- The training dataset has labels or Targets.
- It is used for classification and regression problems.

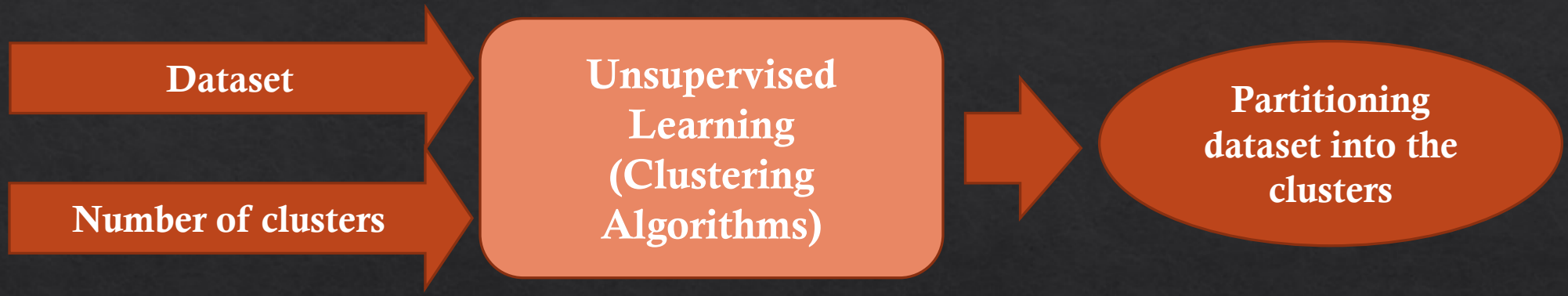
Unsupervised Learning

- It is performed in just one step.
- It deals with unlabelled dataset.
- It is used for clustering problems.

What is the Unsupervised Learning?

- Unsupervised learning is a machine learning technique, where you do not need to supervise the model.
- Instead, you need to allow the model to work on its own to discover information. It mainly deals with the unlabelled data.
- The goal in such unsupervised learning problems may be to discover groups of similar examples within the data, where it is called **clustering**.
- Clustering is an important concept when it comes to unsupervised learning.
- It mainly deals with finding a structure or pattern in a collection of uncategorized data.
- Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data.
- You can also modify how many clusters your algorithms should identify.

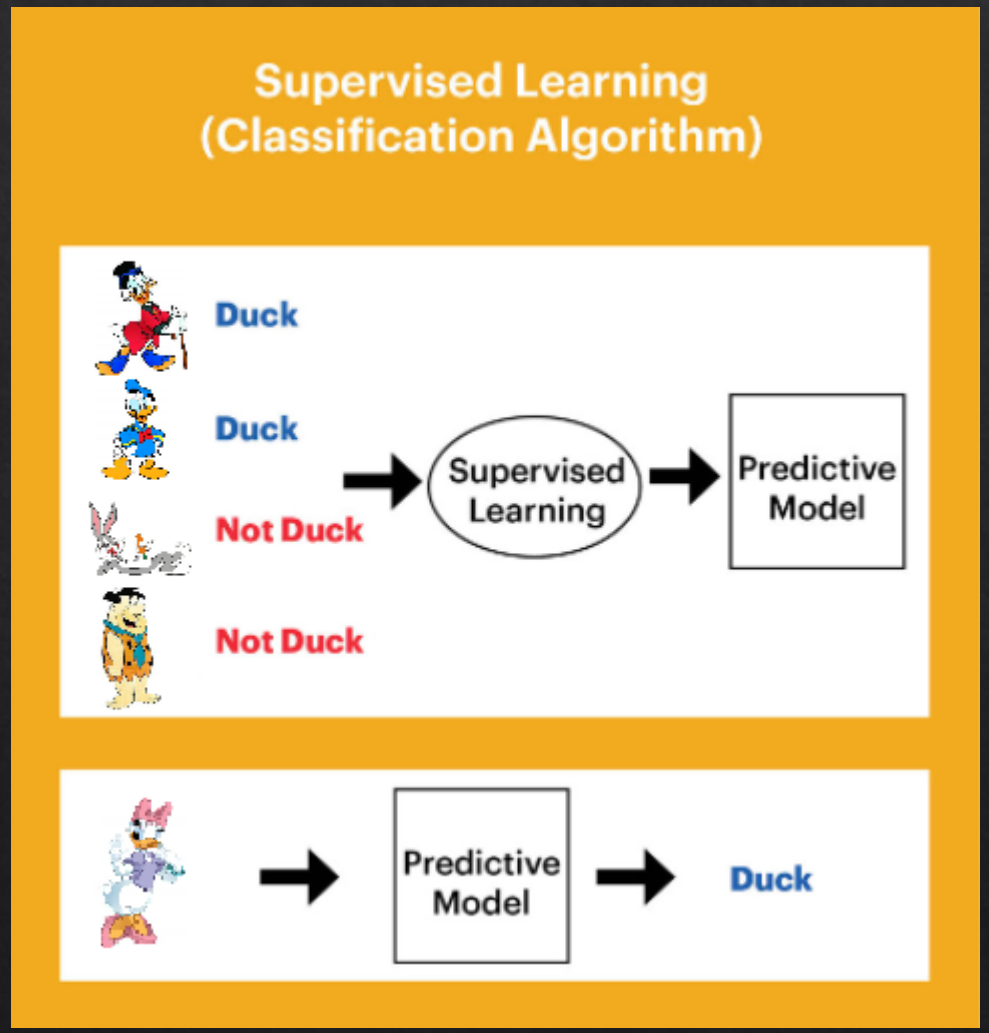
What is the Unsupervised Learning?

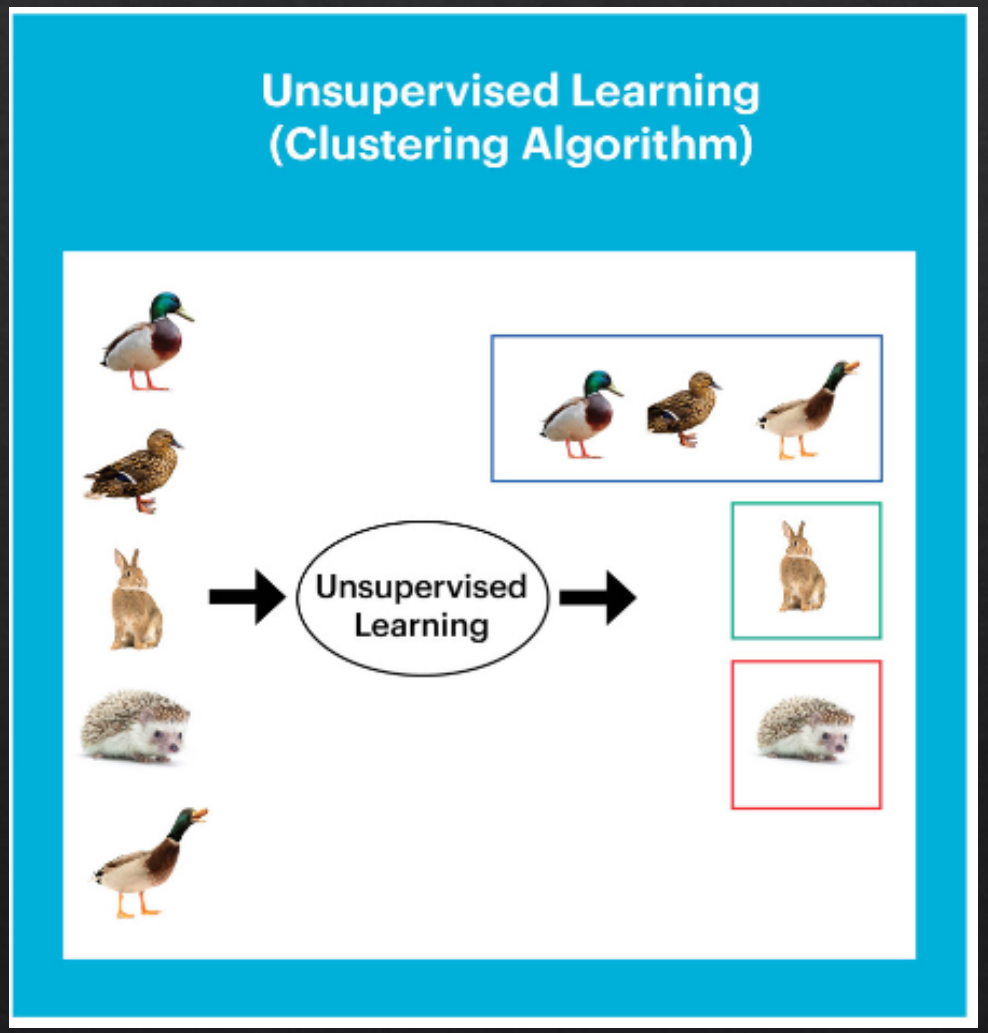


sample

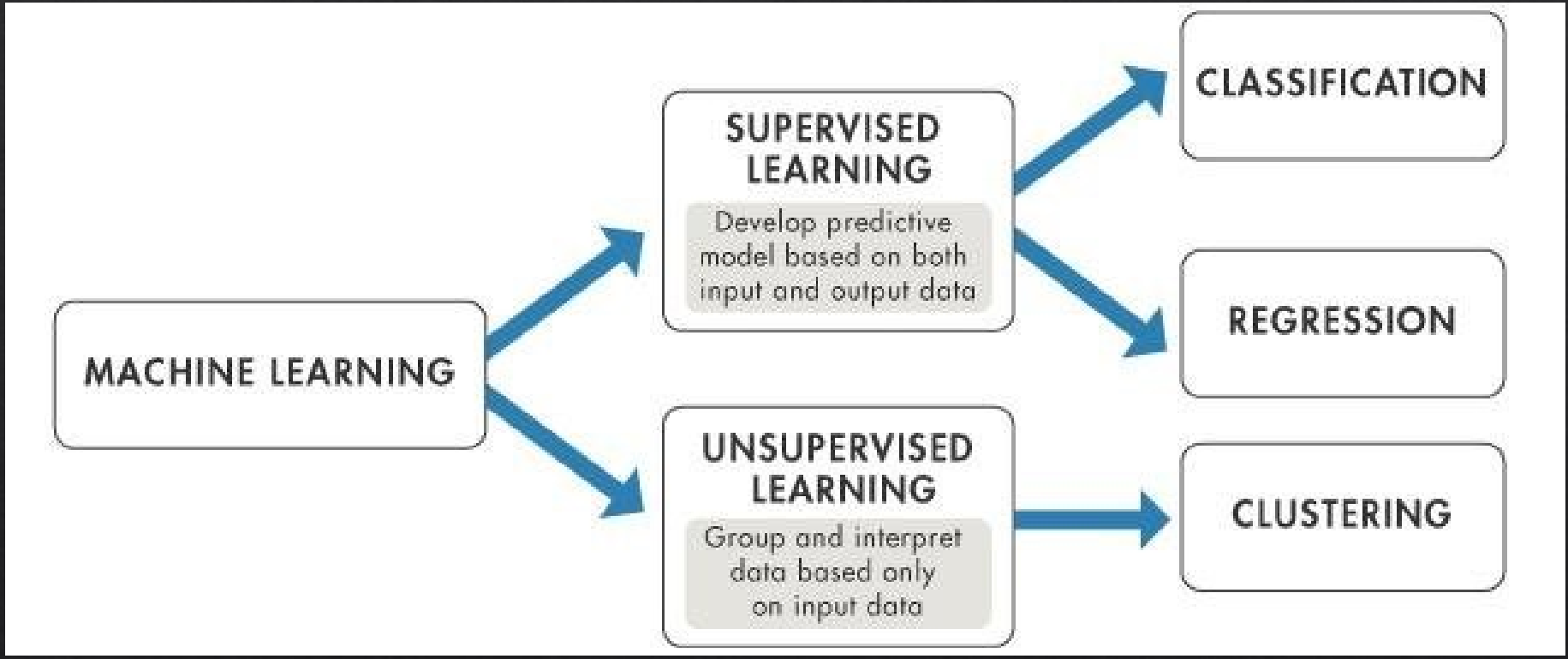
Cluster/group

Some Visual Examples

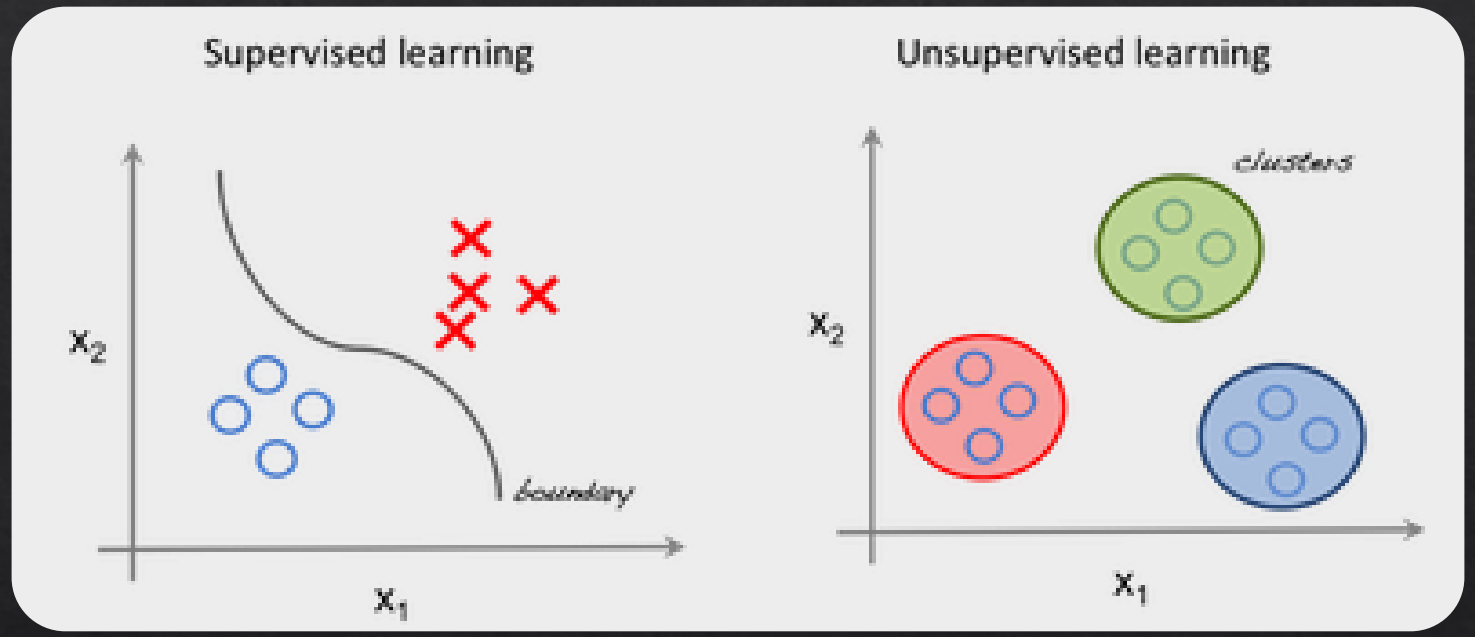




Some Visual Examples

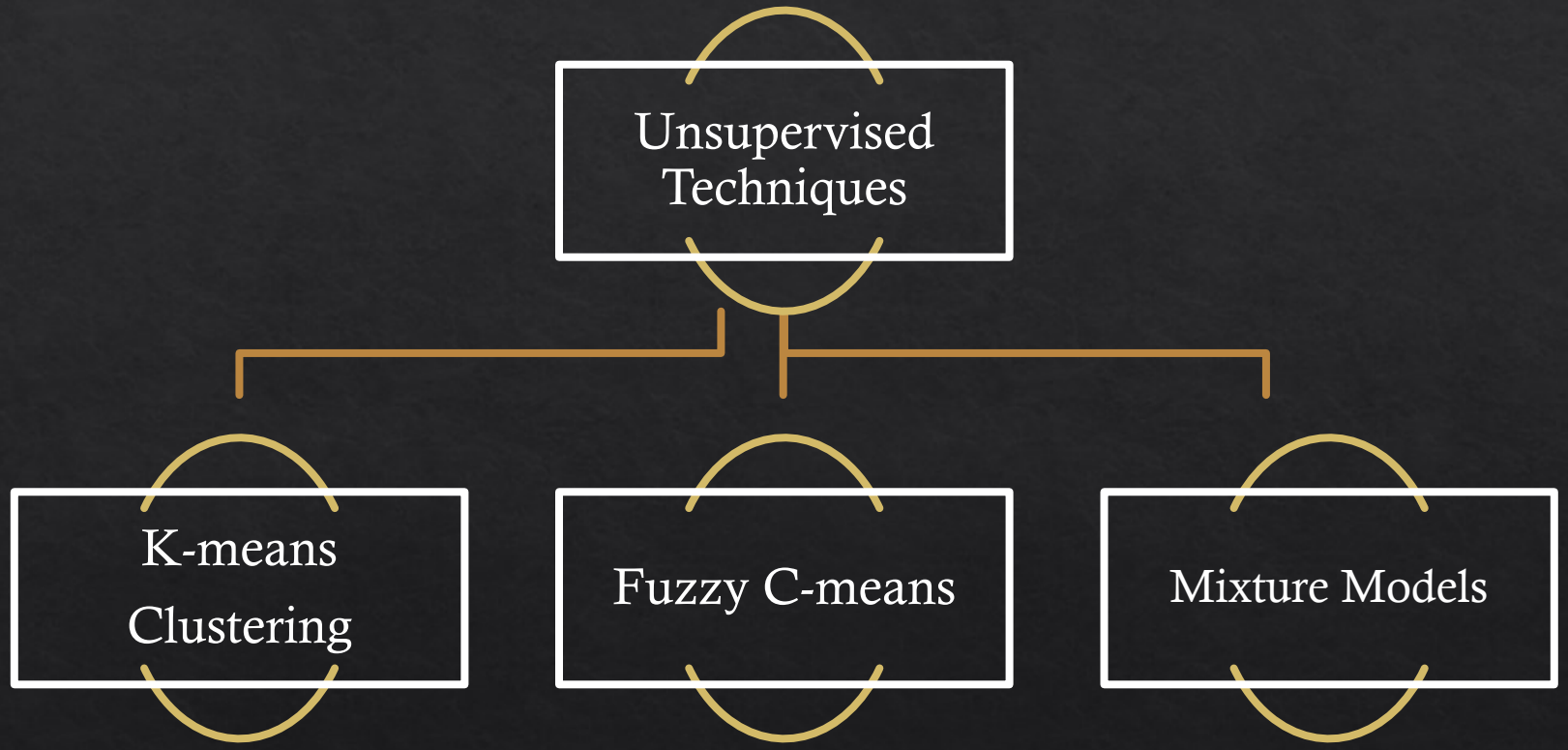


Some Visual Examples



Useful Link:
https://www.guru99.com/supervised-vs-unsupervised-learning.html

Unsupervised Techniques



K-means Clustering

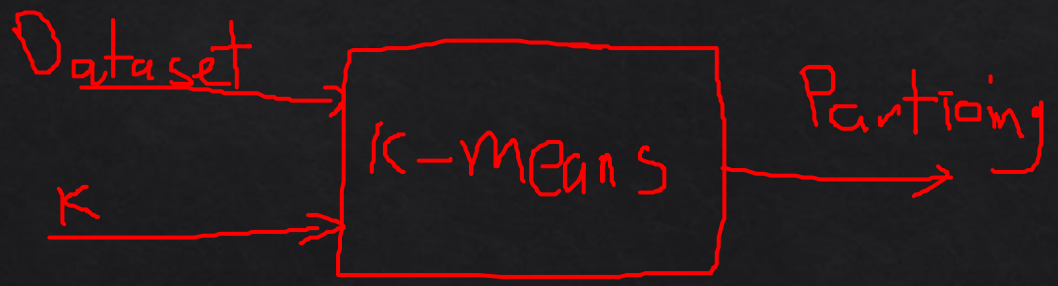
- K-means clustering is one of the simplest and popular unsupervised machine learning algorithms.
- It aims to partition n observations into k clusters using an iterative algorithm.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), where each observation is a d-dimensional real vector, k-means clustering aims to partition the n observations into k ($\leq n$) sets $\mathbf{S} = \{S_1, S_2, ..., S_k\}$ so as to minimize the within-cluster sum of squares (WCSS) (i.e. variance). Formally, the objective is to find:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \arg\min_{\mathbf{S}} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i$$

where $\boldsymbol{\mu}_i$ is the mean of points in S_i. This is equivalent to minimizing the pairwise squared deviations of points in the same cluster:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \frac{1}{2|S_i|} \sum_{\mathbf{x},\mathbf{y} \in S_i} \|\mathbf{x} - \mathbf{y}\|^2$$



K-means Clustering

K-means algorithm:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e assignment of data points to clusters isn't changing.
    - Compute the sum of the squared distance between data points and all centroids.
    - Assign each data point to the closest cluster (centroid).
    - Compute the centroids for the clusters by taking the average of the all data points that belong to each cluster.

K-means Clustering