UNIVERSITY OF PORTSMOUTH

Python for Data Analysis

Modeling in Python - Clustering

(TB2 - Week 5)

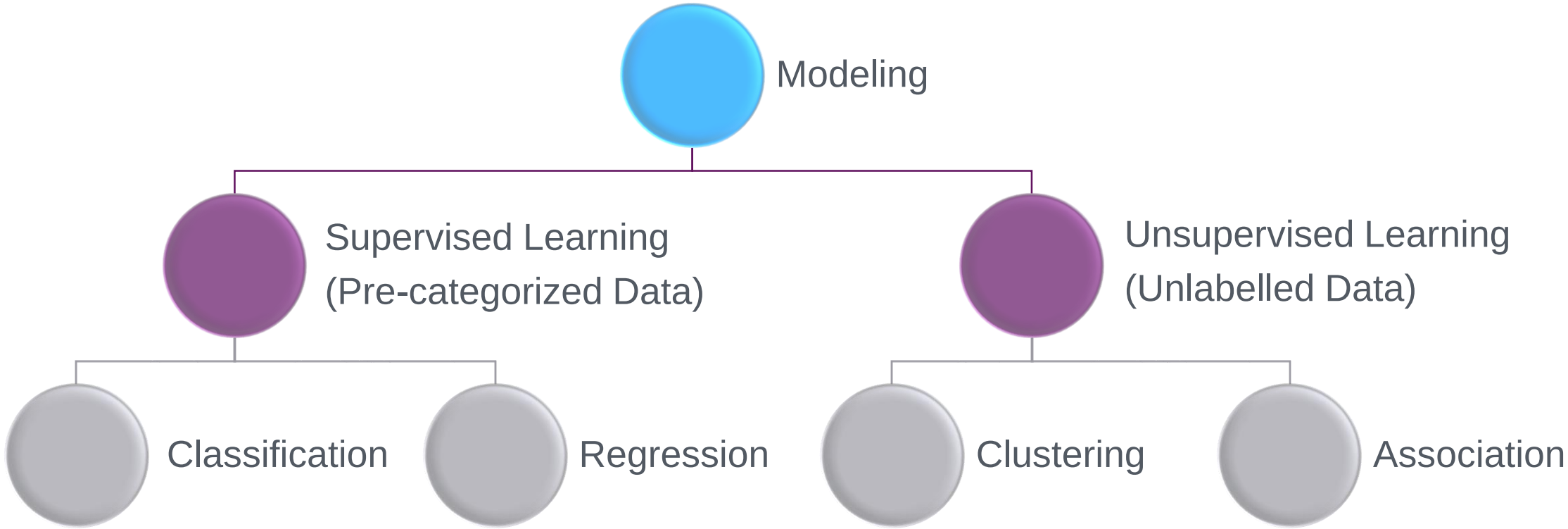
Atefeh Khazaei

atefeh.khazaei@port.ac.uk

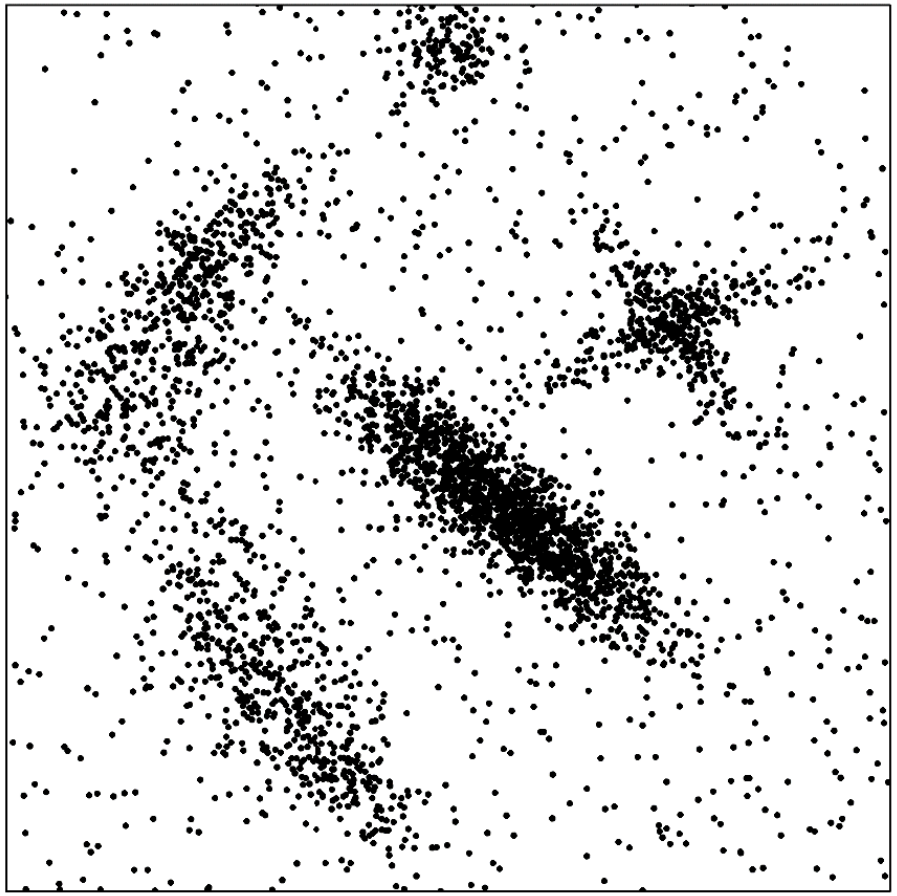
What we will learn this week?

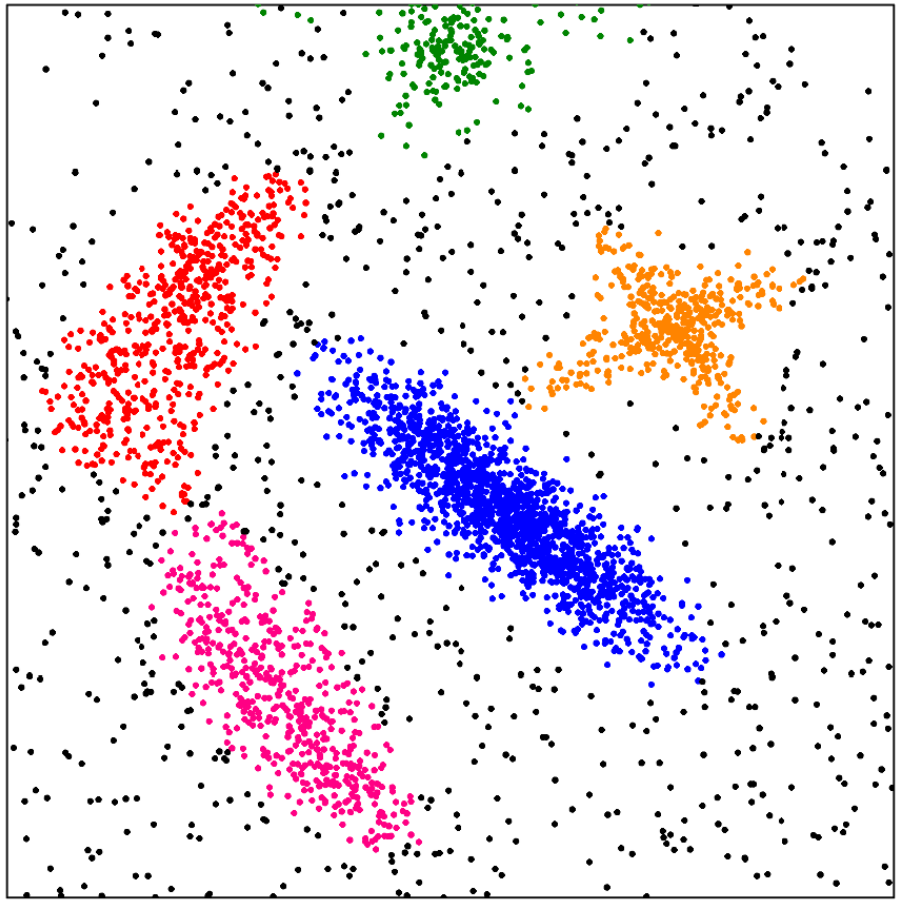
- Supervised vs. Unsupervised Learning
- Introduction to Clustering
  - K-Means

Supervised vs. Unsupervised Learning



What is Cluster Analysis?





What is Cluster Analysis? (cont.)

- **Cluster**: A collection of data objects
  - Similar (or related) to one another within the same group: High intra-class similarity
  - Dissimilar (or unrelated) to the objects in other groups: Low inter-class similarity
- Cluster analysis (or clustering, data segmentation, …)
  - **Finding similarities between data according to the characteristics found in the data and grouping similar data objects into clusters**
- Typical applications
  - As a stand-alone tool to get insight into data distribution
  - As a pre-processing step for other algorithms

Partitioning Clustering Algorithms:
Basic Concept

- The **simplest and most fundamental** version of cluster analysis
- Organizes the objects of a set into several exclusive groups or clusters.
- The clusters are formed to **optimize an objective partitioning criterion**.
- A partitioning criterion within cluster variation: Partitioning a database *D* of *N* objects into a set of *k* clusters, such that **the sum of squared distances is minimized** (where c_i is the centroid or medoid of cluster C_i)

$$E = \Sigma_{i=1}^{k} \Sigma_{p \in C_i} (d(p, c_i))^2$$

K-means Algorithms

Algorithm: *k*-means. The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

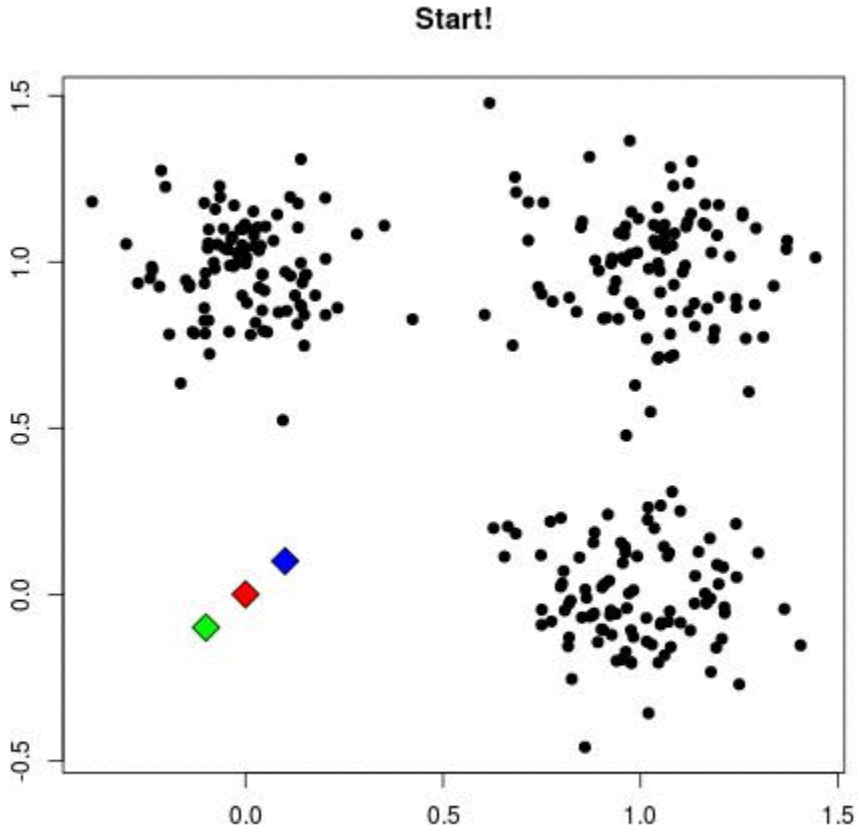
Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

Iterative algorithm

- **Initialization**
- **Iterative body**
- **Stop condition**

K-means Algorithms (cont.)



Animated GIF Ref: https://towardsdatascience.com/cluster-analysis-create-visualize-and-interpret-customer-segments-474e55d00ebb

sklearn.cluster.KMeans()

- You can find more details related to k-means in the following link:
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

- Different parameters of k-means
- More examples

Clustering Evaluation
Error Sum of Squares (SSE)

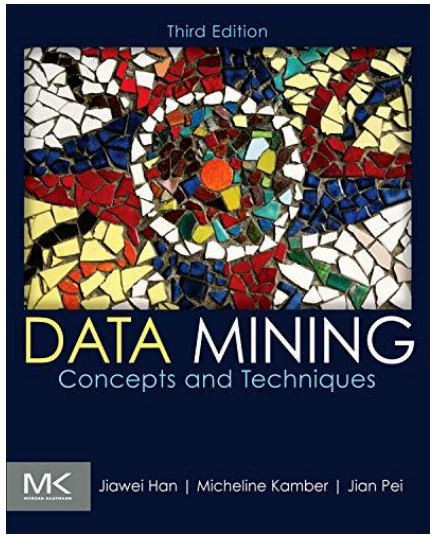
- **Error Sum of Squares (SSE)** is the sum of the squared differences between each observation (sample) and its group's mean (centroid).

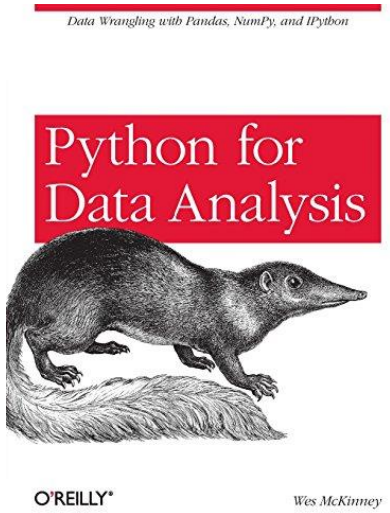
$$SSE = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- It can be used as a measure of variation within a cluster.
- The **lower the SSE is more desirable** and it means that the more similar samples are in each cluster.

References & More Resources

- References:
  - McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and Ipython*, O'Reilly Media, Inc., 2012.
  - Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.





Practical Session

- Please download TB2_Week05_Kmeans.ipynb file, and run it to learn new points.
- Read more details related to k-means and its parameters in the following link:
  - https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- Build some other clustering k-means models.
- Try different parameters for these models and compare them together.