# Lecture 5

## Clustering and Classification

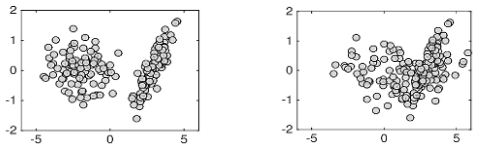
## Objectives

At the end of the session, you should be able to

1. differentiate between supervised and unsupervised learning; and
2. perform a clustering algorithm in Python.

## Discovering Patterns from Data

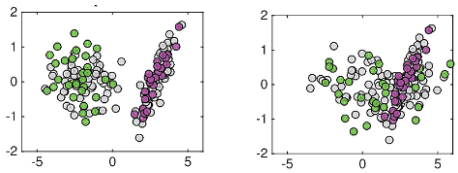
**Image Source:** Data-driven science and engineering : machine learning, dynamical systems, and control/ Steven L. Brunton



If you were tasked to group these data points how would you do the grouping?

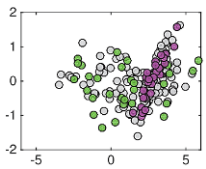
## Discovering Patterns from Data

**Image Source:** Data-driven science and engineering : machine learning, dynamical systems, and control/ Steven L. Brunton



If you were tasked to group these data points how would you do the grouping?

## Supervised vs Unsupervised Learning



In supervised machine learning, the algorithm is presented with labelled datasets. Thus examples of the input and output of a desired model are explicitly given.

**Input**

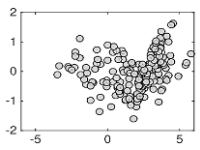
data $\{\mathbf{x}_j \in \mathbb{R}^n,\ j \in Z := \{1, 2, \cdots, m\}\}$

labels $\{\mathbf{y}_j \in \{\pm 1\},\ j \in Z' \subset Z\}$

**Output**

labels $\{\mathbf{y}_j \in \{\pm 1\},\ j \in Z\}$.

## Supervised vs Unsupervised Learning



No labels are given for unsupervised learning algorithms. They must find patterns in the data in a principled way.

**Input**

data $\{\mathbf{x}_j \in \mathbb{R}^n,\ j \in Z := \{1, 2, \cdots, m\}\}$

**Output**

labels $\{\mathbf{y}_j \in \{\pm 1\},\ j \in Z\}$.

## k-Means Clustering

*k*-means clustering is one of the most prominent unsupervised algorithms that is in use today.

*Goal* : Partition a set of vector valued data of $m$ observations into $k$ clusters.

*Idea* : Each observation is labeled as belonging to a cluster with the nearest mean, which serves as a proxy for that cluster.

## k-Means Clustering

Algorithm

(i) Given initial values for $k$ distinct means, compute the distance of each observation $\mathbf{x}_j$ to each of the $k$ means.

(ii) Label each observation as belonging to the nearest mean.

(iii) Once labeling is completed, find the center-of-mass (mean) for each group of labeled points.

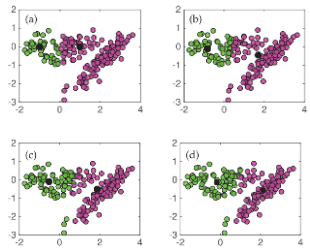
(iv) If stopping criterion is not met, return to step (i) with the updated $k$ means.

## k-Means Clustering

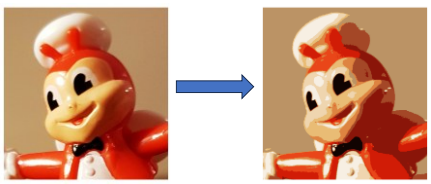
Possible Stopping Criteria

1. Centroids of newly formed clusters do not change
2. Points remain in the same cluster
3. Maximum number of iterations is reached

## Illustration: k-Means Clustering




## Application: Color quantization



## END