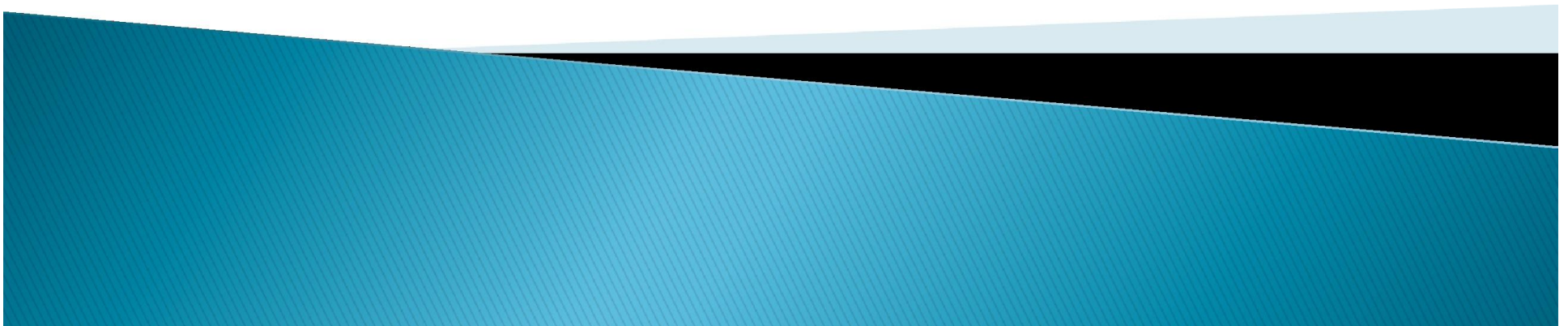


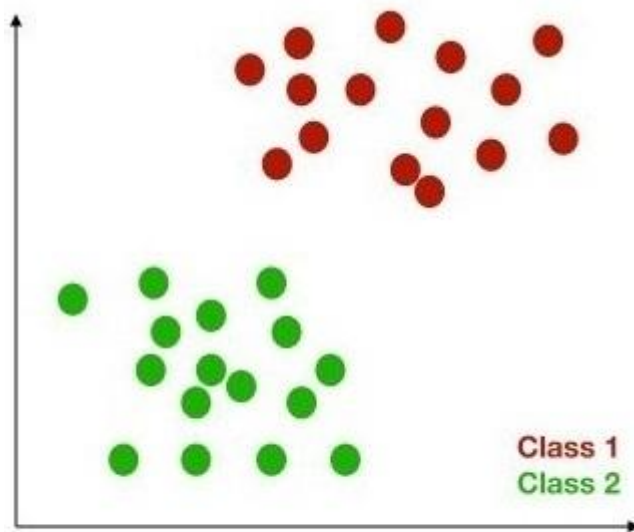
# Unsupervised Learning

**Nazerfard, Ehsan**  
nazerfard@aut.ac.ir



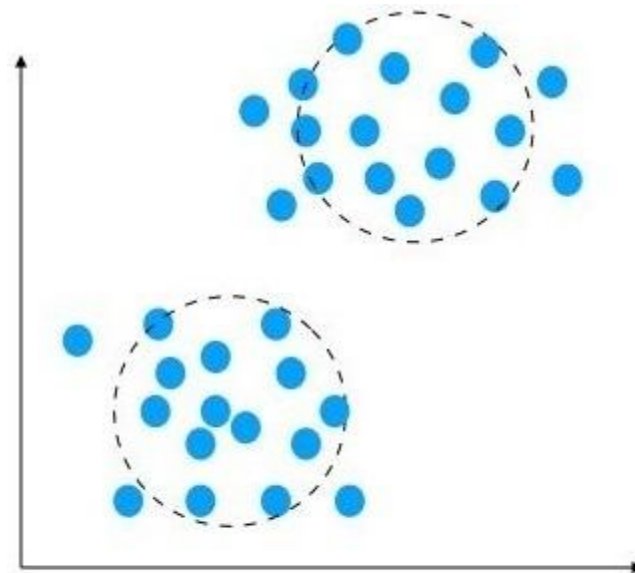
# Supervised vs. Unsupervised

## □ Predictive Modeling



Supervised

## □ Descriptive Modeling



Unsupervised

$$D1_{train} = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2) \dots, (\mathbf{X}_m, y_m)\}$$

$$D2_{train} = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_m\}$$

# Unsupervised Learning

- ❑ In supervised learning, we are given features together with targets (i.e. class labels).
- ❑ In **unsupervised learning**, we are only given features.
  - The task is to discover structures in the data.
- ❑ Why discover structures?

# Unsupervised Learning (cont.)

- ❑ In supervised learning, we are given features together with targets (i.e. class labels).
- ❑ In **unsupervised learning**, we are only given features.
  - The task is to discover structures in the data.
- ❑ Why discover structures?
  - More meaningful representation for the data
    - Dimensionality reduction
    - Density Estimation
    - Clustering
    - Pre-processing technique prior to applying other ML models
    - ...

# Dimensionality Reduction

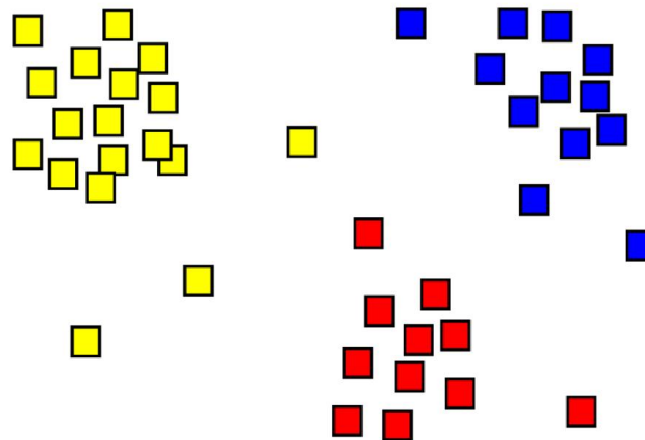
- ❑ The curse of dimensionality
- ❑ Dimensionality reduction is simply the process of reducing the dimension of your feature set.
  - ❑ Dimension reduction methods come in **supervised** and **unsupervised** forms.
- ❑ **Supervised** methods include linear discriminant analysis (LDA), which is designed to find low-dimensional projection that maximizes class separation.
- ❑ **Unsupervised** methods include principal components analysis (PCA), which rotates and projects data along the direction of increasing variance. The features with the maximum variance are the principal components.

# Density Estimation

- Estimating continuous probability density function from observed data
- Imagine that you have some data points  $x_1, \dots, x_n$  that come from common unknown distribution  $f$ .
  - Histogram (the most basic estimate)
  - Parzen window or kernel density estimation:
    - Approximate  $f$  by a mixture of continuous distributions, called kernels, that are centered at  $x_i$  data points.

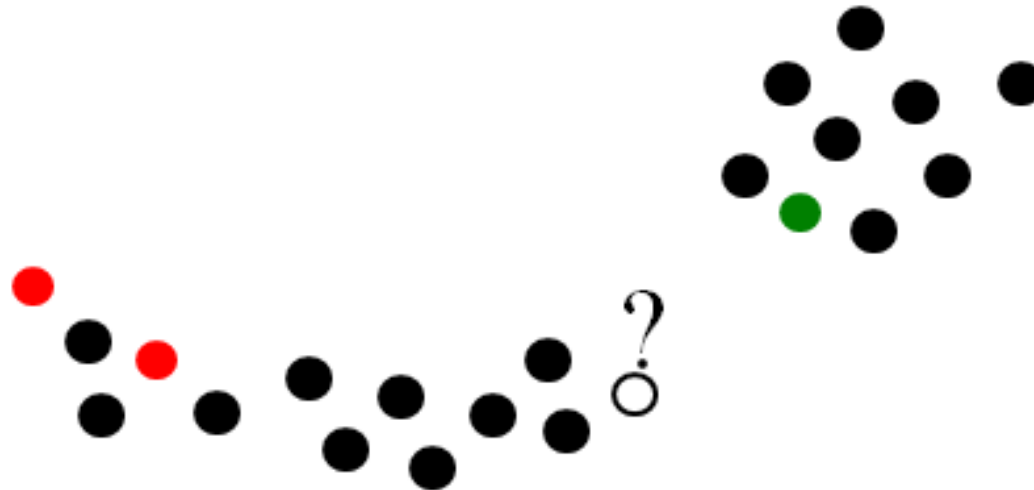
# Clustering

- Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (in some sense) to each other than to those in other groups.



# Pre-processing

- ❑ Clustering can be used in process of semi-supervised Learning. It can be used first to find natural segmentation of the data and then create labels.
  - Crowd sourcing
- ❑ The pre-processed data with labels can then be used to develop a semi-supervised classification.





# Further Reading (cont.)

## □ Semi-supervised learning

- Lots of data is available, but only small portion is labeled.
  - J van Engelen, H Hoos, [A survey on semi-supervised learning](#), Machine Learning, 109:373–440, 2020.

## □ Transfer Learning

- When labeled data is scarce for a specific target task, transfer learning often offers an effective solution by utilizing data from a related source task.
- When negative transfer may occur?
  - W Zhang, L Deng, L Zhang, D Wu, [Overcoming Negative Transfer: A Survey](#), IEEE Trans. On Knowledge and Data Engineering, 2020.

