Unsupervised models #1

Madson L. D. Dias e Lucas S. Sousa

Huawei / IFCE

March 8, 2021

Agenda

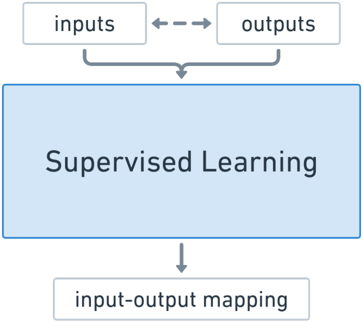
1. Recap
2. Clustering
   - Connectivity-based clustering
   - Centroid-based clustering
   - Distribution-based clustering
   - Density-based clustering
   - Clustering using k-means
3. Dimensionality reduction
   - Principal Component Analysis

Agenda

1. Recap
2. Clustering
   - Connectivity-based clustering
   - Centroid-based clustering
   - Distribution-based clustering
   - Density-based clustering
   - Clustering using k-means
3. Dimensionality reduction
   - Principal Component Analysis

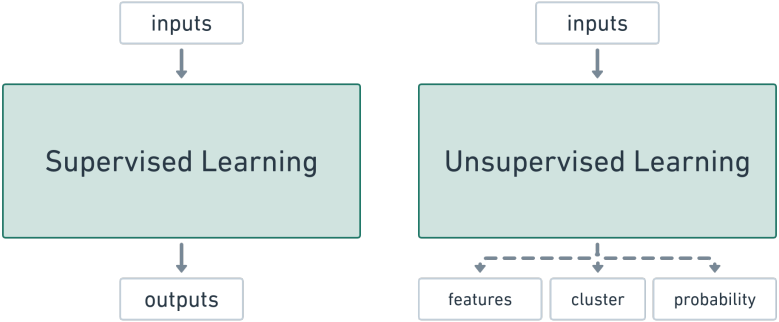
Recap on Unsupervised learning

- A type of machine learning algorithm used to draw inferences from data sets consisting of input data without labeled responses
- Task of inferring a function to describe **hidden structure** from unlabeled data.



Recap on Unsupervised learning

- A type of machine learning algorithm used to draw inferences from data sets consisting of input data without labeled responses
- Task of inferring a function to describe **hidden structure** from unlabeled data.



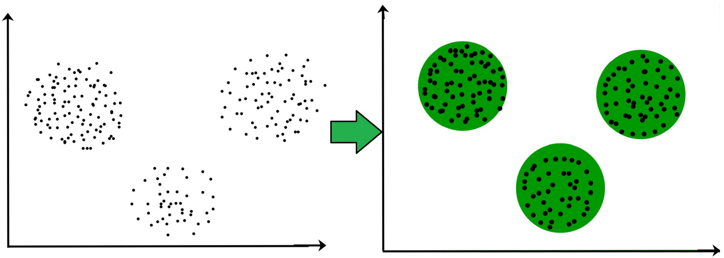
Recap on Unsupervised learning

Some applications

Recap on Unsupervised learning

Some applications

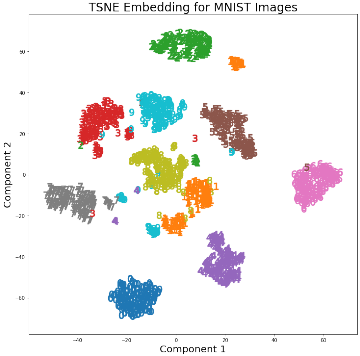
- **Clustering**
- Dimensionality reduction
- Anomaly detection
- ...



Recap on Unsupervised learning

Some applications

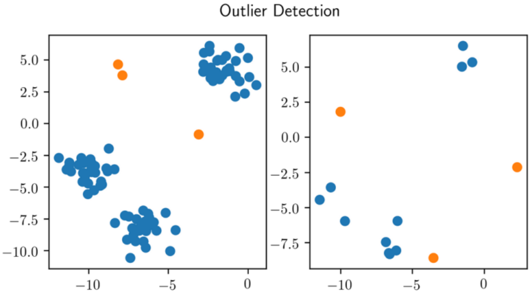
- Clustering
- **Dimensionality reduction**
- Anomaly detection
- ...



Recap on Unsupervised learning

Some applications

- Clustering
- Dimensionality reduction
- **Anomaly detection**
- ...



Agenda

1. Recap
2. Clustering
   - Connectivity-based clustering
   - Centroid-based clustering
   - Distribution-based clustering
   - Density-based clustering
   - Clustering using k-means
3. Dimensionality reduction
   - Principal Component Analysis

Clustering

- Given a dataset $\mathcal{D} = \{\boldsymbol{x}_n \in \mathbb{R}^D\}_{n=1}^N$, where $\boldsymbol{x}$ is a vector with D dimensions and N is the number of elements of the data set[1]. The task of clustering is to find a function $g(\boldsymbol{x})$ in that $g : \mathbb{R}^D \to \mathbb{N}^K$, where $\mathcal{C}$ is a set of clusters $\mathcal{C} = \{\mathcal{C}_k\}_{k=1}^{K<N}$, and $\mathcal{C}_k = \{\boldsymbol{x}_n : g(\boldsymbol{x}_n) = k\}$ such that

$$\mathcal{C}_k \neq \emptyset, \forall k \tag{1}$$

$$\bigcup_{k=1}^{K} \mathcal{C}_k = \mathcal{D} \tag{2}$$

$$\mathcal{C}_k \bigcap \mathcal{C}_j = \emptyset,\ \forall k, j \text{ where } k \neq j \tag{3}$$

[1] The $\mathbb{R}^D$ space is called "data space" or "input space".

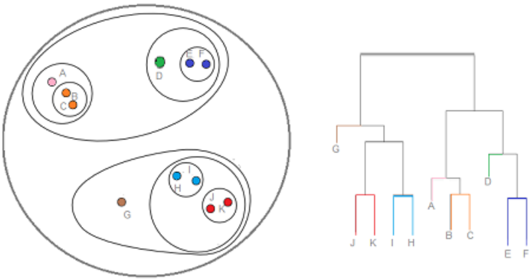
Clustering

- 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 (clusters).
- Categorization of clustering algorithms
  - $\rightarrow$ Connectivity-based clustering
  - $\rightarrow$ Centroid-based clustering
  - $\rightarrow$ Distribution-based clustering
  - $\rightarrow$ Density-based clustering
  - $\rightarrow$ Grid-based clustering

Agenda

Connectivity-based clustering

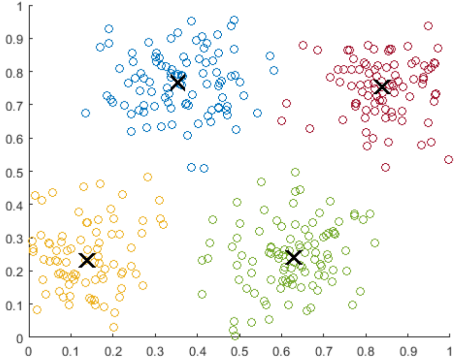
- Also known as hierarchical clustering, is based on the core idea of objects being more related to nearby objects than to objects farther away.



Agenda

Centroid-based clustering

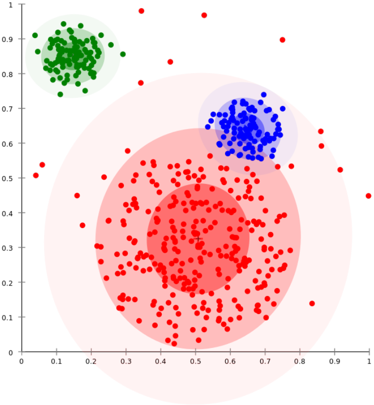
- The clusters are represented by a central vector (centroid), which may not necessarily be a member of the data set.



Agenda

Distribution-based clustering

- Clusters can then easily be defined as objects belonging most likely to the same distribution.

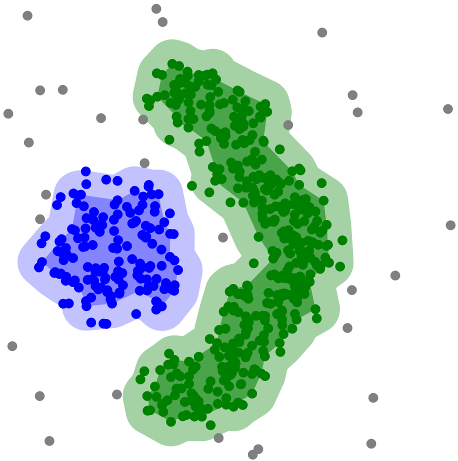


Agenda

1. Recap
2. Clustering
   - Connectivity-based clustering
   - Centroid-based clustering
   - Distribution-based clustering
   - Density-based clustering
   - Clustering using k-means
3. Dimensionality reduction
   - Principal Component Analysis

Density-based clustering

- Clusters are defined as areas of higher density than the remainder of the data set.



Agenda

k-means clustering

Definition

- Given a dataset $\mathcal{D} = \{\boldsymbol{x}_n \in \mathcal{R}^D\}_{n=1}^N$, k-means clustering aims to partition the N observations into $K(\leq N)$ sets so as to minimize the *within-cluster* sum of squares (i.e. variance). Formally, the objective is to find:

$$\mathcal{C}^\star = \arg\min_{\mathcal{C}} \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in \mathcal{C}_k} \|\boldsymbol{x} - \boldsymbol{\mu}_k\|^2 \tag{4}$$

where $\boldsymbol{\mu}_k$ is the centroid of the cluster $\mathcal{C}_k$.

k-means clustering

Algorithm

Algorithm 1 k-means algorithm

Specify the number k of clusters to assign.
Randomly initialize k centroids.
repeat
    expectation: Assign each point to its closest centroid.
    maximization: Compute the new centroid (mean) of each cluster.
until The centroid positions do not change.

"Talk is cheap.
Show me the code."

- Linus Torvalds

Agenda

Dimensionality reduction

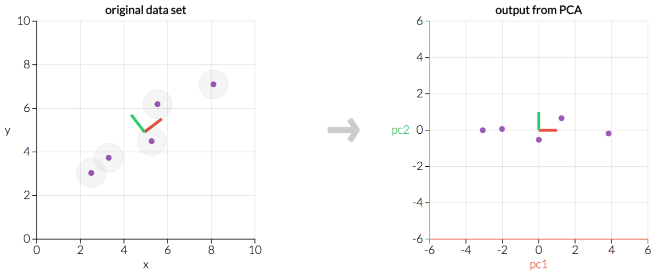
- Transformation of data from a high-dimensional space into a low-dimensional space so that the low-dimensional representation retains some meaningful properties of the original data, ideally close to its *intrinsic dimension*.
- Methods
  - $\rightarrow$ Feature selection: filter strategy, the wrapper strategy, and the embedded strategy
  - $\rightarrow$ Feature projection: Principal component analysis (PCA), ..., Autoencoder

Agenda

1. Recap
2. Clustering
   - Connectivity-based clustering
   - Centroid-based clustering
   - Distribution-based clustering
   - Density-based clustering
   - Clustering using k-means
3. Dimensionality reduction
   - Principal Component Analysis

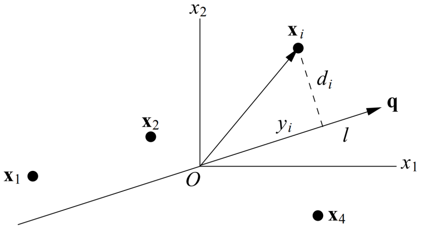
Principal Component Analysis

- PCA is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.



Principal Component Analysis

- Consider a set of points $\mathcal{D} = \{\boldsymbol{x}_n\}_{n=1}^{N}$ in a D-dimensional space, such that their mean $\boldsymbol{\mu} = \mathbf{0}$, i.e., centroid is at the origin.



- The PCA want to finds a line l through the origin that maximizes the projections y_n of the points $\boldsymbol{x}_n$ on l.
- Let $\boldsymbol{q}$ denote the unit vector along line l.

Principal Component Analysis

- The projection y_n of $\boldsymbol{x}_n$ on l is $y_n = \boldsymbol{x}_n^\top \boldsymbol{q}$.
- The mean squared projection is the variance V over all points

$$
\begin{aligned}
V &= \frac{1}{N}\sum_{n=1}^{N} y_n^2 && (5)\\
&= \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}_n^\top \boldsymbol{q})^2 && (6)\\
&= \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{q}^\top \boldsymbol{x}_n)(\boldsymbol{x}_n^\top \boldsymbol{q}) && (7)\\
&= \boldsymbol{q}^\top \left[\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top\right] \boldsymbol{q} && (8)
\end{aligned}
$$

Principal Component Analysis

- The middle factor is the covariance matrix $\mathbf{C}$ of the data points

$$\mathbf{C} = \frac{1}{N} \sum_{n=1}^{N} \boldsymbol{x}_n \boldsymbol{x}_n^\top \tag{9}$$

- We want to find a unit vector $\boldsymbol{q}$ that maximizes the variance V

$$\boldsymbol{q}^\star = \arg\max_{\boldsymbol{q}} \boldsymbol{q}^\top \mathbf{C} \boldsymbol{q}, \tag{10}$$

- Put $\|\boldsymbol{q}\| = 1$ constraint to avoid overflow
- The constraint optimization problem

$$\text{maximize}\, V = \boldsymbol{q}^\top \mathbf{C} \boldsymbol{q} \quad \text{subject to} \quad \|\boldsymbol{q}\| = 1. \tag{11}$$

Lagrange multiplier method

- Consider this problem

$$\text{maximize} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad g(\boldsymbol{x}) = c. \tag{12}$$

- Lagrange multiplier method introduces a Lagrange multiplier λ to combine $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ as

$$L(\boldsymbol{x}, \lambda) = f(\boldsymbol{x}) - \lambda(g(\boldsymbol{x}) - c) \tag{13}$$

- Then, we can solve using

$$\frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \boldsymbol{x}} = 0, \quad \frac{\partial L(\boldsymbol{x}, \lambda)}{\partial \lambda} = 0 \tag{14}$$

Principal Component Analysis

- Use Lagrange multiplier method, combine V and the constraint
- Consider this problem

$$\text{maximize} \quad L = \boldsymbol{q}^\top \mathbf{C} \boldsymbol{q} - \lambda(\boldsymbol{q}^\top \boldsymbol{q} - 1) \tag{15}$$

- Now, we differentiate L with respect to $\boldsymbol{q}$ and λ:

$$\frac{\partial L}{\partial \boldsymbol{x}} = 2\boldsymbol{q}^\top \mathbf{C} - 2\lambda \boldsymbol{q}^\top = 0 \tag{16}$$

$$\frac{\partial L}{\partial \lambda} = \boldsymbol{q}^\top \boldsymbol{q} - 1 = 0 \tag{17}$$

- Eq. (16) gives

$$\boldsymbol{q}^\top \mathbf{C} = \lambda \boldsymbol{q}^\top \iff \mathbf{C} \boldsymbol{q} = \lambda \boldsymbol{q} \tag{18}$$

- This is called an *eigenvector equation*

Principal Component Analysis

General PCA

- PCA transforms $\boldsymbol{x}_n$ into a new vector $\boldsymbol{y}_n$ through Q as follows:

$$\boldsymbol{y}_n = \mathbf{Q}^\top(\boldsymbol{x}_n - \boldsymbol{\mu}) = \sum_{m=1}^{M} (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{q}_m \boldsymbol{q}_m \tag{19}$$

- Each component of $\boldsymbol{y}_n$ is

$$y_{nm} = (\boldsymbol{x}_n - \boldsymbol{\mu})^\top \boldsymbol{q}_m \tag{20}$$

- This is the projection of $\boldsymbol{x}_n - \boldsymbol{\mu}$ on $\boldsymbol{q}_m$.

Thanks for your attention!