UNIVERSIDAD NACIONAL
DE SAN MARTÍN

Machine Learning - Week 3

Maestría en Ciencia con mención de Tecnología de la información

Yonatan Carlos CARRANZA ALARCÓN,
Ph. D. in Machine Learning,
ycarranza.alarcon@gmail.com

https://salmuz.github.io/

October 16, 2021

Overview

Overview

Supervised vs. Unsupervised learning [Patil, 2018].

Figure: Supervised learning

- Labeled dataset $(\boldsymbol{x}_i, y_i)_{i=1}^N$.
- Learn a decision boundary model.

Supervised vs. Unsupervised learning [Patil, 2018].

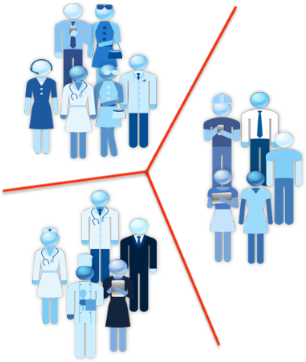


Figure: Supervised learning

Figure: Unsupervised learning

- Labeled dataset $(\boldsymbol{x}_i, y_i)_{i=1}^{N}$.
- Learn a decision boundary model.

- Unlabeled dataset $(\boldsymbol{x}_i)_{i=1}^{N}$.
- Discover hidden patterns in dataset without human intervention.

Supervised vs. Unsupervised learning [Patil, 2018].

Figure: Supervised learning

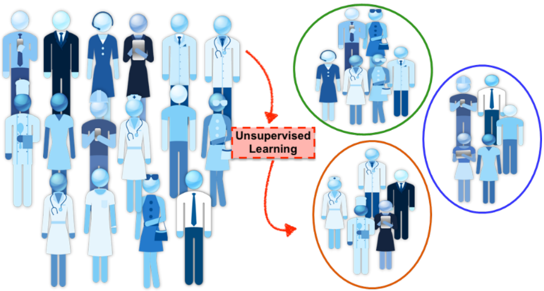


Figure: Unsupervised learning

- Labeled dataset $(\boldsymbol{x}_i, y_i)_{i=1}^{N}$.
- Learn a decision boundary model.

- Unlabeled dataset $(\boldsymbol{x}_i)_{i=1}^{N}$.
- Discover hidden patterns in dataset without human intervention.

Unsupervised learning - Motivation.

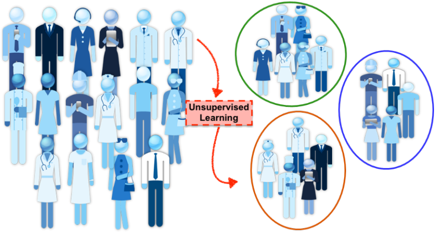


Figure: Unsupervised learning

- Unlabeled dataset $(\boldsymbol{x}_i)_{i=1}^{N}$.
- Discover hidden patterns in dataset without human intervention.

Why would you use unlabeled data?

- To label data is expensive in time and money (e.g. Biology).
- Often labeled data is not available.
- In BigData, it is difficult properly to label all data (e.g. Crowdsourcing).

Unsupervised learning.

Unsupervised learning is a type of machine learning in which the algorithm is not provided with any pre-assigned labels or scores for the training data [Hinton et al., 1999].

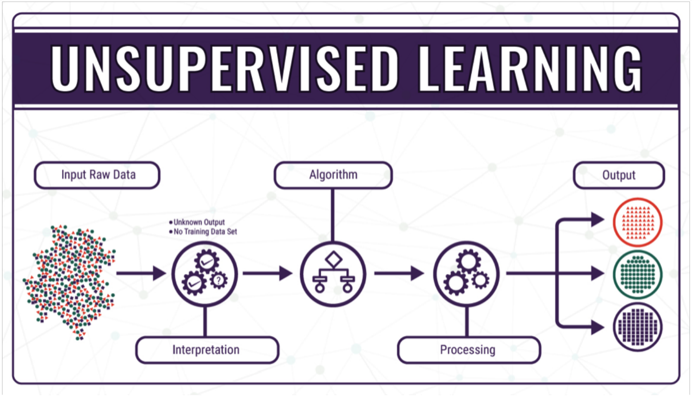


Figure: Outline of Unsupervised learning scheme

Unsupervised learning.

Objective

It may be to discover groups of similar examples within the data (i.e. clustering),

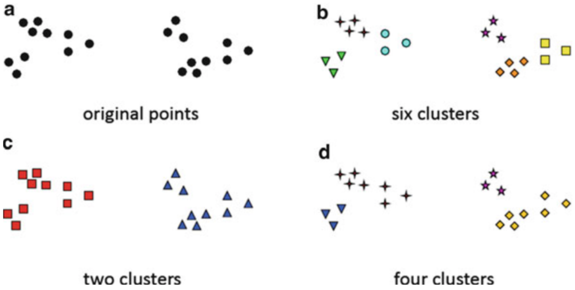


Figure: Different ways of clustering the same set of points [Dougherty, 2012].

Unsupervised learning.

Objective

It may be to discover groups of similar examples within the data (i.e. clustering),

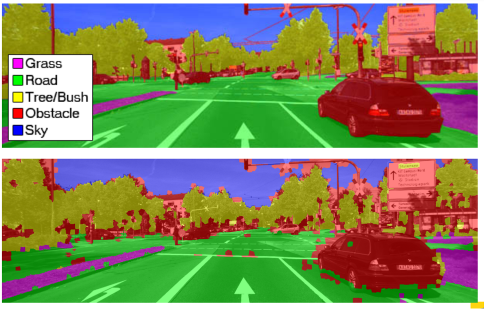


Figure: U-Net architecture for image segmentation Ronneberger et al. [2015].

Unsupervised learning.

Objective

It may be to discover groups of similar examples within the data (i.e. clustering), or to determine the distribution of data within the input space (i.e. density estimation),

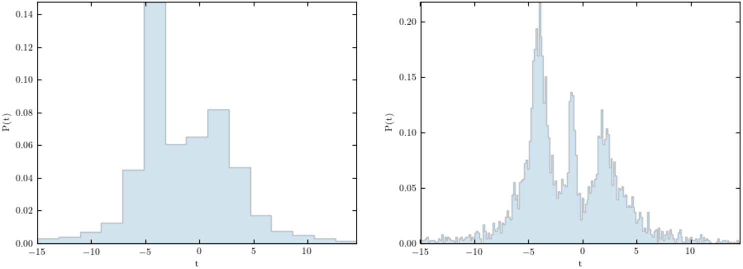


Figure: Density estimation from a raw data set.

Unsupervised learning.

Objective

It may be to discover groups of similar examples within the data (i.e. clustering), or to determine the distribution of data within the input space (i.e. density estimation), or to project the data from a high-dimensional space down to two or three dimensions for the purpose of visualization.

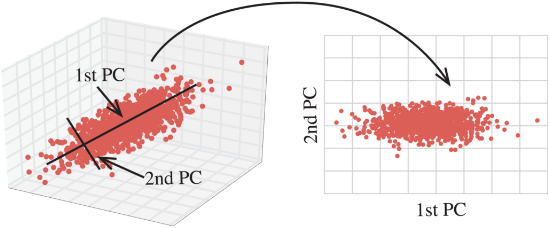


Figure: Principal Component Analysis for dimensionality reduction

Mathematical formulation

Let $\mathscr{D} = \{\boldsymbol{x}_i | i = 1, \ldots, N\} \subseteq \mathscr{X}^p$ be a data set generated from an unknown joint probability distribution $\mathbb{P}_X$

- A vector of p predictors $\boldsymbol{x}$ (also called inputs, features, attributes, explanatory variables)

Objective

The goal is to build a function $\varphi : \mathscr{X} \to \mathcal{O}$ that finds or discovers hidden and interesting patterns (in an $\mathcal{O}$ output space) in unlabeled data by minimizing or maximizing a specific criterion $\mathcal{C} : \mathscr{X}^{\otimes N} \to \mathbb{R}$.

For instances: Learning labeled from raw data.

There are several algorithmic, statistical, and mathematical methods!

Overview

Clustering

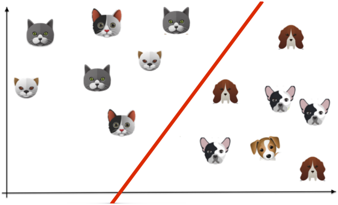


Figure: Labeled samples

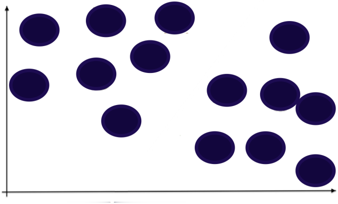


Figure: Unlabeled samples

- How can we discover hidden patterns ?
- How can we measure the dissimilarity between two samples?

Clustering

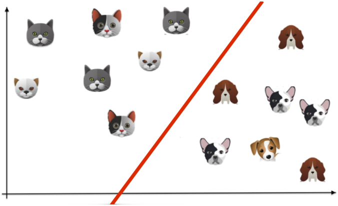


Figure: Labeled samples

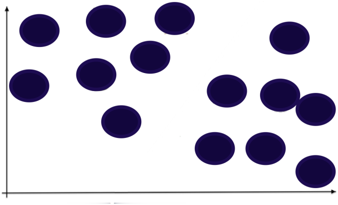


Figure: Unlabeled samples

- How can we discover hidden patterns ?
- How can we measure the dissimilarity between two samples?

Dissimilarity measure

It correspond to the intuitive idea of a distance between two objects: the larger it is, the farther the objects are.

$$D(x_i, x_j) = \text{dist}(x_i, x_j), \quad i, j \in \{1, \ldots, N\} \tag{1}$$

Clustering

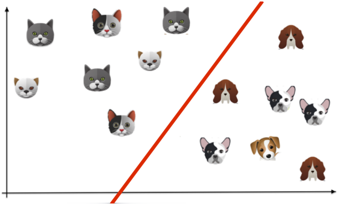


Figure: Labeled samples

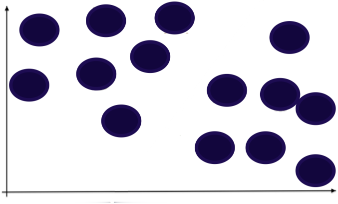


Figure: Unlabeled samples

$$D(x_i, x_j) = \mathrm{dist}(x_i, x_j), \quad i, j \in \{1, \dots, N\}$$

- Which dissimilarity measure or topological space?
  $\rightarrow$ For practical purposes, we use an euclidien topological space!

Clustering

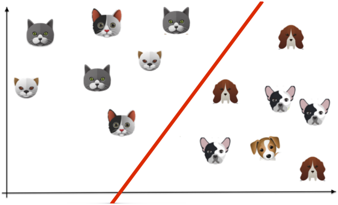


Figure: Labeled samples

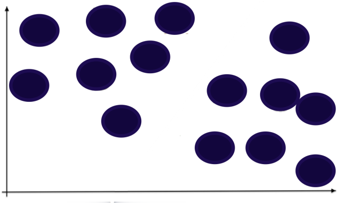


Figure: Unlabeled samples

- Which dissimilarity measure or topological space?
  $\rightarrow$ For practical purposes, we use an euclidien topological space!

Euclidean distance

It is the distance between the two points in n-dimensional Euclidean space:

$$D(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2, \quad i, j \in \{1, \ldots, N\} \tag{1}$$

K-means clustering

The K-means algorithm is one of the most popular iterative descent clustering methods used often for variables of quantitative type Friedman et al. [2001].

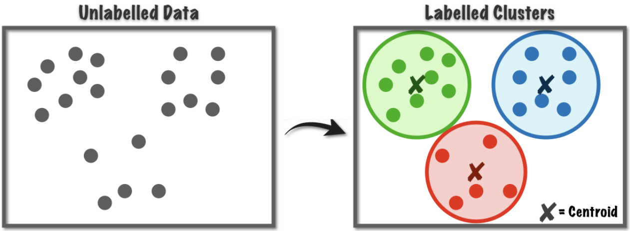
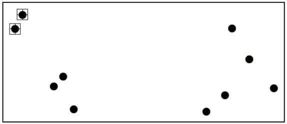
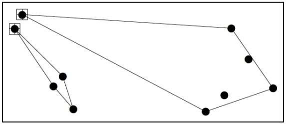


Figure: K-means clustering example

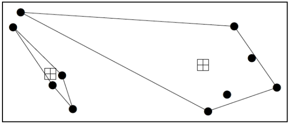
K-means clustering (Algorithm)



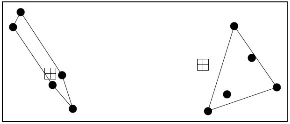
(a) Step 1: Choice 2 randoms points



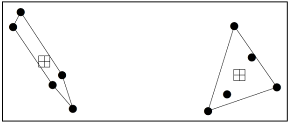
(b) Step 2: Assignment of each point to the nearest center



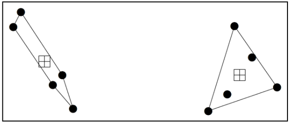
(c) Step 3: Calculate the new centers of gravity.



(d) Step 4: Assignment of each point to the nearest center

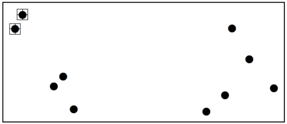


(e) Step 5: Calculate the new centers of gravity.

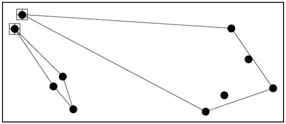


(f) Step 6: Assignment of each point to the nearest center

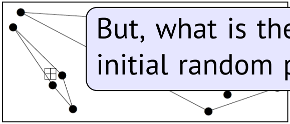
K-means clustering (Algorithm)



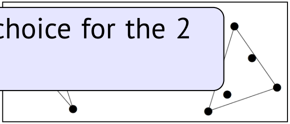
(a) Step 1: Choice 2 randoms points



(b) Step 2: Assignment of each point to the nearest center

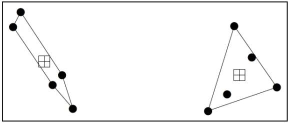


(c) Step 3: Calculate the new centers of gravity.

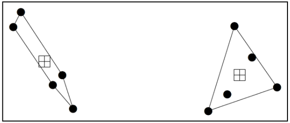


(d) Step 4: Assignment of each point to the nearest center

But, what is the better choice for the 2 initial random points?



(e) Step 5: Calculate the new centers of gravity.



(f) Step 6: Assignment of each point to the nearest center

K-means clustering (Quality of a partition)

How can we obtain good quality of a clustering partition?

Inertia and variance

Inertia measure the dispersion of observations relative to a reference point u in a metric space.

$$I_u(\{\xi_i, w_i\}_{i=1}^n) = \sum_{i=1}^n w_i \|\xi_i - u\|_M = \sum_{i=1}^n w_i \|x_i\|_M = I_O(\{x_i, w_i\}_{i=1}^n)$$

where $w_i = 1/n$ and M is often choice as the matrix identity.

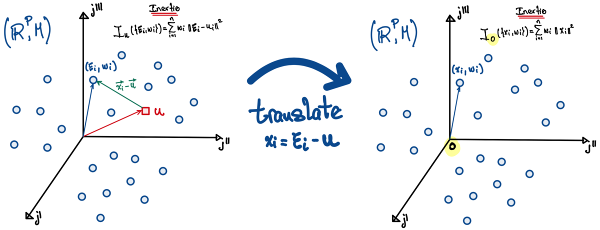


Figure: Translation of observations into u as origine.

K-means clustering (Quality of a partition)

How can we obtain good quality of a clustering partition?

Huygens' theorem

Total inertia $I_O(\{x_i, w_i\}_{i=1}^n)$ of points on a partition $P = (P_1, \ldots, P_K)$ can be written as follows

Total inertia ofthe points = Inertia Intra-cluster + Inertia Inter-cluster.

$$I_O(\{x_i, w_i\}_{i=1}^n) = \sum_{k=1}^{K} w^k \|\bar{x}_k\|_M^2 + \sum_{k=1}^{K} \sum_{i \in P_k} w_i \|x_i - \bar{x}_k\|_M^2, \qquad \text{(Inertie totale)}$$

$$\sum_{k=1}^{K} w^k \|\bar{x}_k\|_M^2, \quad \text{where:} \quad w^k = \sum_{i \in P_k} w_i \qquad \text{(Inertia Intra-cluster)}$$

$$\sum_{k=1}^{K} \sum_{i \in P_k} w_i \|x_i - \bar{x}_k\|_M^2 = \sum_{k=1}^{K} I_{\bar{x}_k}(P_k), \qquad \text{(Inertia Inter-cluster)}$$

K-means clustering (Quality of a partition)

Geometrical Insights of Huygens' theorem

$$I_O(\{x_i, w_i\}_{i=1}^n)$$

$$\sum_{k=1}^{K} w^k \|\bar{x}_k\|_M^2 \quad + \quad \sum_{k=1}^{K} \sum_{i \in P_k} w_i \|x_i - \bar{x}_k\|_M^2$$

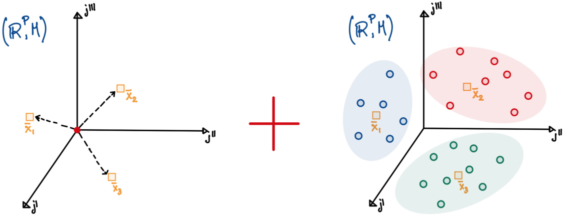


Figure: Huygens' theorem

K-means clustering (Quality of a partition)

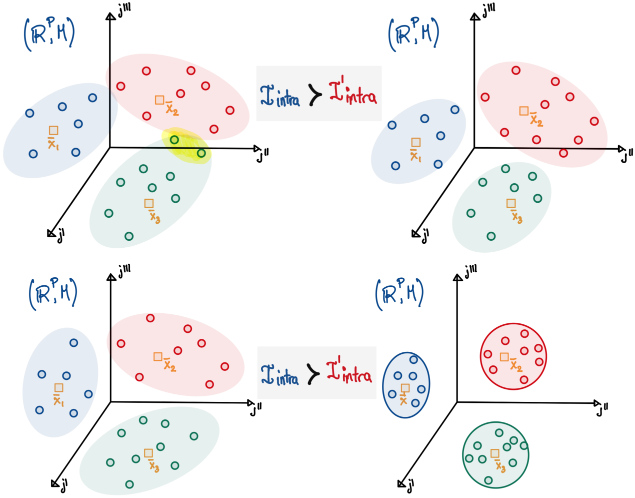


Figure: What is the better partition?

Hierarchical clustering

Let us consider N observations (or objects) $\mathcal{O} = \{o_1, \ldots, o_N\}$ and a dissimilarity measure between pairs of observations $d(\cdot, \cdot)$ on $\mathcal{O}$.

In Hierarchical clustering, we also define, based on $d(\cdot, \cdot)$, a linkage criterion $\mathcal{C}(\cdot, \cdot)$ which specifies the dissimilarity of sets as a function of the pairwise distances of observations in the sets.

Let us consider in this course tree linkage criterions:

- Minimum or single-linkage clustering

$$\mathcal{C}(A, B) = \min \left\{ d(\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{x} \in A \text{ and } \boldsymbol{x}' \in B, A, B \subseteq \mathcal{O} \right\}$$

- Maximum or complete-linkage clustering

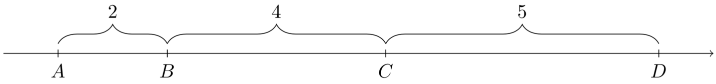
$$\mathcal{C}(A, B) = \max \left\{ d(\boldsymbol{x}, \boldsymbol{x}'), \boldsymbol{x} \in A \text{ and } \boldsymbol{x}' \in B, A, B \subseteq \mathcal{O} \right\}$$

- Unweighted average linkage clustering

$$\mathcal{C}(A, B) = \frac{1}{n_A * n_B} \sum_{\boldsymbol{x} \in A, \boldsymbol{x}' \in B} d(\boldsymbol{x}, \boldsymbol{x}')$$

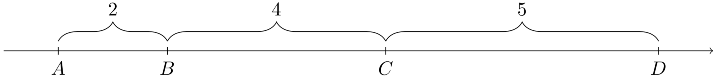
Hierarchical clustering

Let us consider 4 points (or objects) separated by some distance



Hierarchical clustering

Let us consider 4 points (or objects) separated by some distance



	A	B	C	D
A	0			
B	$\boxed{2}$	0		
C	6	4	0	
D	11	9	5	0

	$\{A, B\}$	C	D
$\{A, B\}$	0		
C	$\boxed{4}$	0	
D	9	5	0

	$\{A, B, C\}$	D
$\{A, B, C\}$	0	
D	$\boxed{5}$	0

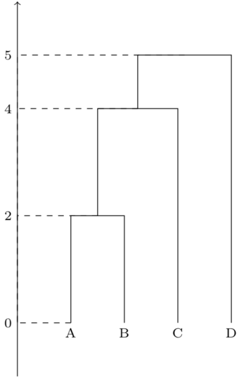
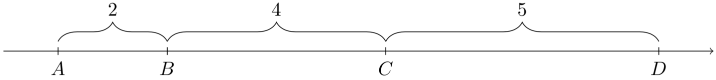


Figure: Minimum or single-linkage clustering

Hierarchical clustering

Let us consider 4 points (or objects) separated by some distance



	A	B	C	D
A	0			
B	$\boxed{2}$	0		
C	6	4	0	
D	11	9	5	0

	$\{A, B\}$	C	D
$\{A, B\}$	0		
C	6	0	
D	11	$\boxed{5}$	0

	$\{A, B\}$	$\{C, D\}$
$\{A, B\}$	0	
$\{C, D\}$	$\boxed{11}$	0

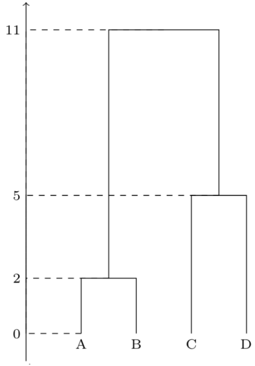
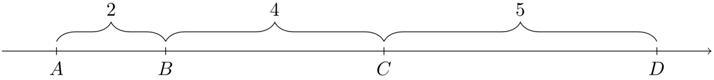


Figure: Maximum or complete-linkage clustering

Hierarchical clustering

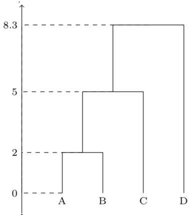
Let us consider 4 points (or objects) separated by some distance



	A	B	C	D
A	0			
B	$\boxed{2}$	0		
C	6	4	0	
D	11	9	5	0

	$\{A, B\}$	C	D
$\{A, B\}$	0		
C	$\boxed{5}$	0	
D	10	$\boxed{5}$	0

	$\{A, B, C\}$	D
$\{A, B, C\}$	0	
D	$\boxed{8.3}$	0



	$\{A, B\}$	$\{C, D\}$
$\{A, B\}$	0	
$\{C, D\}$	$\boxed{7.5}$	0

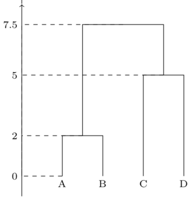
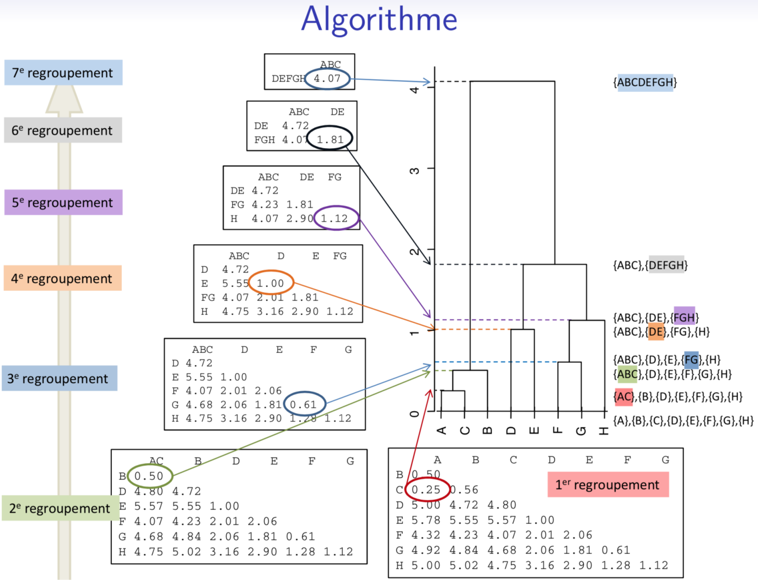


Figure: Unweighted average linkage clustering

Hierarchical clustering (Single-linkage Algorithm)



EM for the Gaussian Mixture Model (GMM)

GMM is a probabilistic clustering assuming that unlabelled observations $\{\boldsymbol{x}_i\}_{i=1}^{N}$ are generated by a mixture of K Gaussian distributions, whose unknown parameters are estimated by an iterative method known as Expectation-Maximization (the EM algorithm).

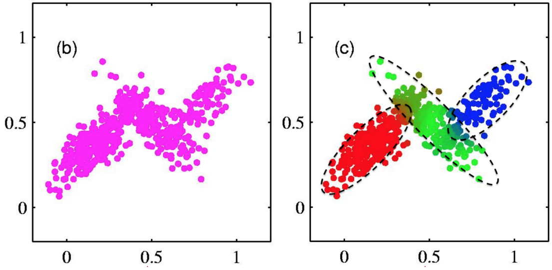


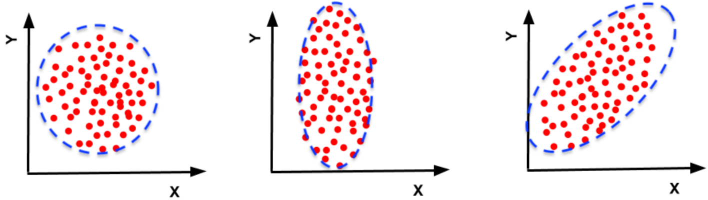
Figure: Examples of Gaussian mixture models

EM for the Gaussian Mixture Model (GMM)

GMM is a probabilistic clustering assuming that unlabelled observations $\{\boldsymbol{x}_i\}_{i=1}^{N}$ are generated by a mixture of K Gaussian distributions, whose unknown parameters are estimated by an iterative method known as Expectation-Maximization (the EM algorithm).

Remark

K-means algorithm is a particular case where all Gaussian distributions are assumed to have **the same diagonal covariance matrix**, with infinitely small variance

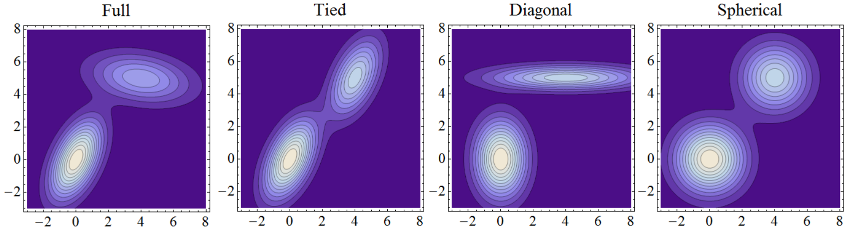


EM for the Gaussian Mixture Model (GMM)

GMM is a probabilistic clustering assuming that unlabelled observations $\{\boldsymbol{x}_i\}_{i=1}^N$ are generated by a mixture of K Gaussian distributions, whose unknown parameters are estimated by an iterative method known as Expectation-Maximization (the EM algorithm).

Remark

K-means algorithm is a particular case where all Gaussian distributions are assumed to have **the same diagonal covariance matrix**, with infinitely small variance



EM for the Gaussian Mixture Model (GMM)

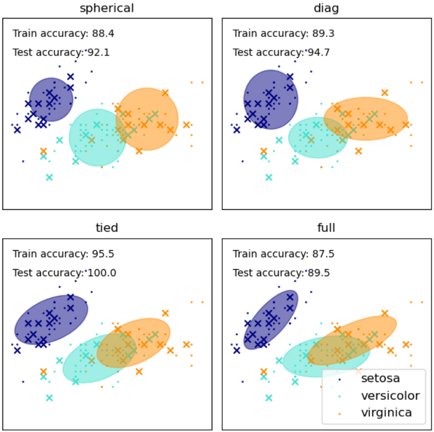


Figure: Gaussian mixture models on **Iris Dataset**

The GaussianMixture comes with different options to constrain the covariance of the difference classes estimated: **spherical, diagonal, tied or full covariance.**

EM Algorithm - Mathematical formulation

- Let $\mathscr{D} = \{\boldsymbol{x}_i\}_{i=1}$ be a unlabelled data set generated i.i.d. from X random variable with probability distribution

$$X \sim \sum_{K}^{k=1} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

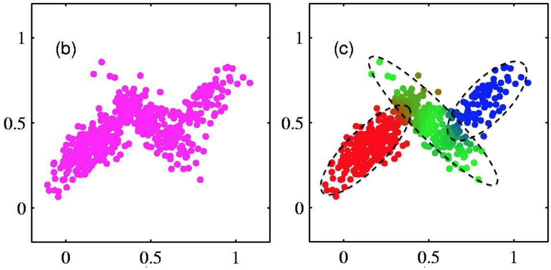


Figure: X follows a Gaussian mixture models

EM Algorithm - Mathematical formulation

- Let $\mathscr{D} = \{\boldsymbol{x}_i\}_{i=1}$ be a unlabelled data set generated i.i.d. from X random variable with probability distribution

$$X \sim \sum_{K}^{k=1} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

- Let's assume that there is a latent random variable $Y \in \{1, \ldots, K\}$ of K classes with $\pi_1, \ldots, \pi_K$ probabilities, such that the condition distribution

$$X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

EM Algorithm - Mathematical formulation

- Let $\mathscr{D} = \{\boldsymbol{x}_i\}_{i=1}$ be a unlabelled data set generated i.i.d. from X random variable with probability distribution

$$X \sim \sum_{K}^{k=1} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$$

- Let's assume that there is a latent random variable $Y \in \{1, \ldots, K\}$ of K classes with $\pi_1, \ldots, \pi_K$ probabilities, such that the condition distribution

$$X|Y = k \sim \mathcal{N}(\mu_k, \Sigma_k)$$

Objective

Estimate the unknown parameters of GMM $\sum_{K}^{k=1} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$

$$\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$$

How to estimate θ? $\Longrightarrow$ Maximum likelihood estimate with EM.

EM Algorithm - Mathematical formulation

Objective

Estimate the unknown parameters of GMM $\sum_{K}^{k=1} \pi_k \mathcal{N}(\mu_k, \Sigma_k)$

$$\theta = (\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K)$$

How to estimate θ? $\Longrightarrow$ Maximum likelihood estimate with EM.

Expectation-Maximization Algorithm

E-Step:

$$\begin{aligned} Q(\theta, \theta^{(t)}) &= \mathbb{E}_{Y|X,\theta^{(t)}} [\log L(\theta; X, Y)] \\ &= \sum_{Y} P(Y|X, \theta^{(t)}) \log(L(\theta; X, Y)) \end{aligned}$$

M-Step:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \theta^{(t)})$$

Loop: $|\ell(\theta^{(t)}; X) - \ell(\theta^{(t+1)}; X)| < \epsilon$

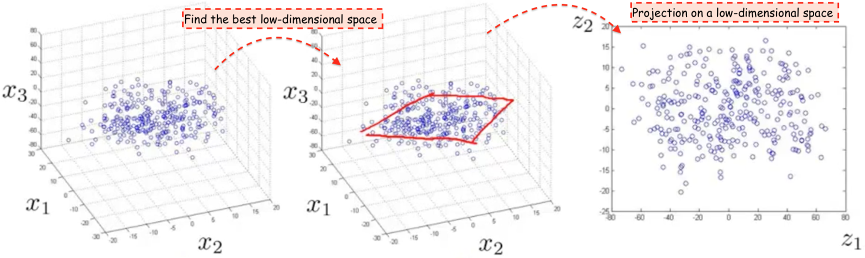
Examples of Clustering

Let us do Machine Learning
Code source - [Link]

Overview

Dimensionality reduction

Dimensionality reduction is the 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.



Dimensionality reduction

Dimensionality reduction is the 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.

Dimensionality reduction - Mathematical formulation

Let $\mathcal{X}^p$ be the original input space of p-dimensional space. Dimensionality reduction aims to find a map function φ^k such that the target output space is of k-dimensional space, with $p > k$, i.e.:

$$\varphi : \mathcal{X}^p \to \mathcal{Y}^k$$

Note that the topological space can be a euclidian space or a manifold space, or whatever other space.

Principal component analysis (PCA)

PCA is a technique for reducing the dimensionality, increasing interpretability but at the same time minimizing information loss.

- Retains geometric properties.
- No losing too much of information or variability.

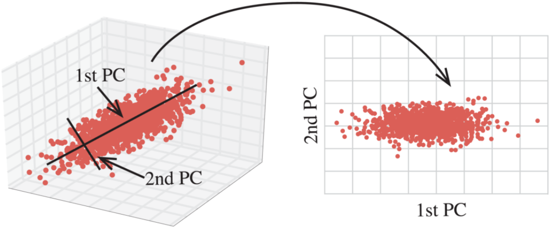


Figure: Example PCA 3D to 2D

Principal component analysis (PCA)

Objective of PCA

PCA aims to:

1. identify hidden pattern in a data set,
2. reduce the dimensionality of the data by removing the noise and redundancy in the data,
3. identify correlated variables

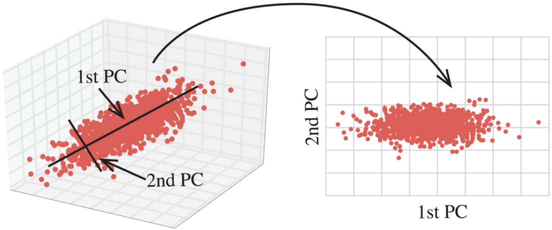


Figure: Example PCA 3D to 2D

Principal component analysis (PCA)

How to minimize the information loss and to maximize the conservation of variability (geometric properties)?

In a Euclidian space, we use the **inertia measure** to calculate the conservation of variability (variance), i.e. the information loss or not.

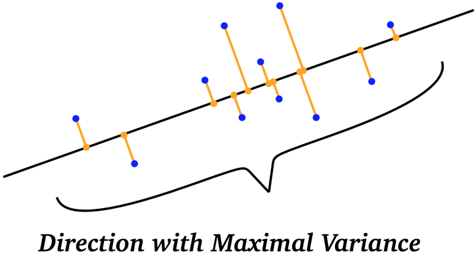


Figure: Maximal Inertia or Variance

Principal component analysis (Inertia measure)

Let be a data set of 20-dimensional space.

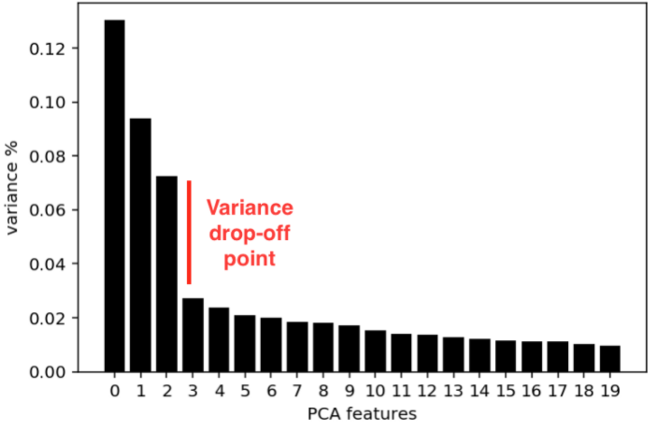


Figure: Variability retains par component

PCA as an identifier of hidden patterns

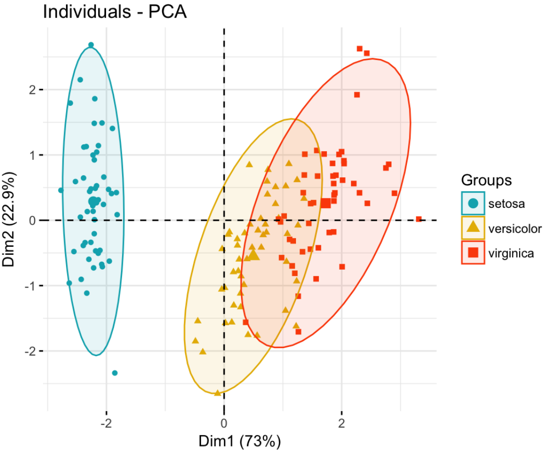


Figure: PCA applied to Iris data set Kassambara et al. [2017].

PCA as an identifier of hidden patterns

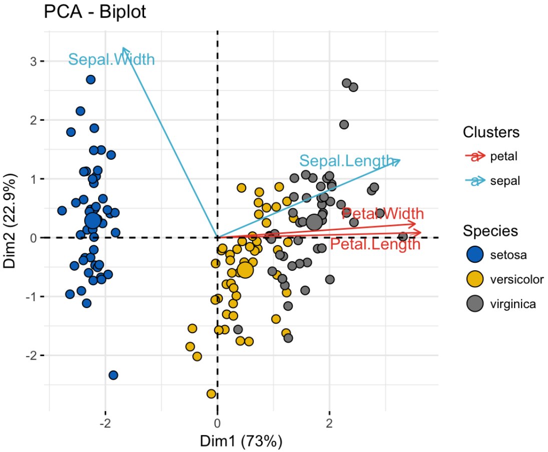


Figure: PCA applied to Iris data set Kassambara et al. [2017].

PCA as an identifier of hidden patterns

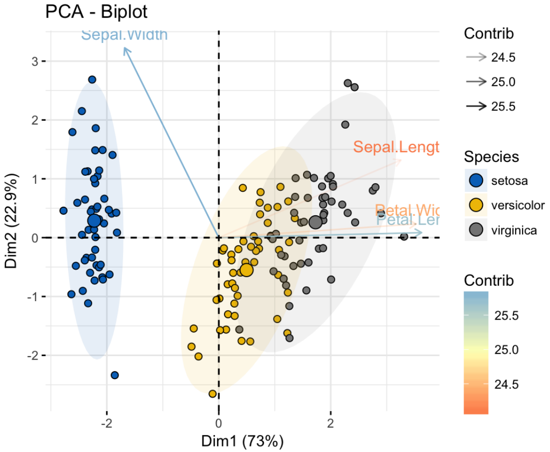


Figure: PCA applied to Iris data set Kassambara et al. [2017].

Principal component analysis (Drawbacks)

PCA sometime does not reduce very well the origin space.

Figure: Which animal is?

Principal component analysis (Drawbacks)

PCA sometime does not reduce very well the origin space.

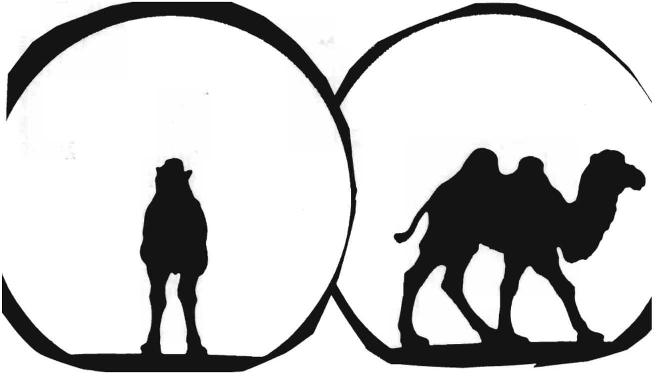


Figure: Which animal is?

Principal component analysis (Drawbacks)

PCA sometime does not reduce very well the origin space.

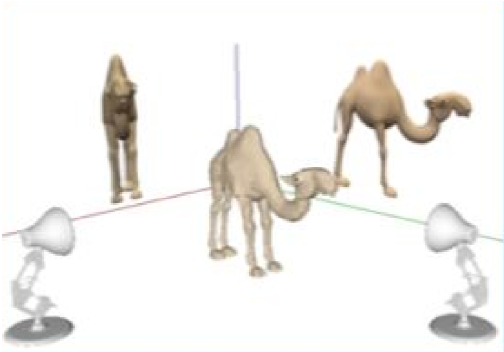


Figure: Which animal is?

t-SNE

PCA is not suitable for non linear data, but T-SNE can work with that.

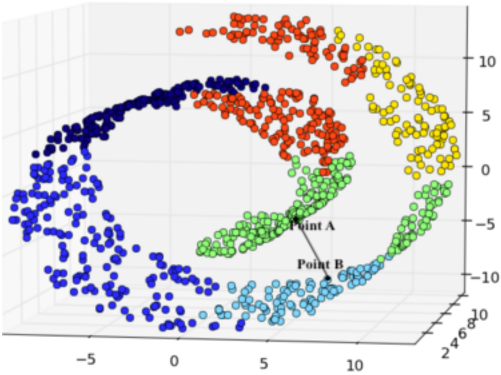


Figure: Non-linear data set

t-SNE

Dimensionality reduction

It is a statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.

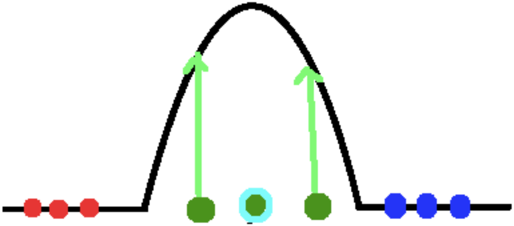
Non-linear data

It is a nonlinear dimensionality reduction technique well-suited for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions

t-SNE

t-SNE algorithm (two stages) Wikipedia [2021b]

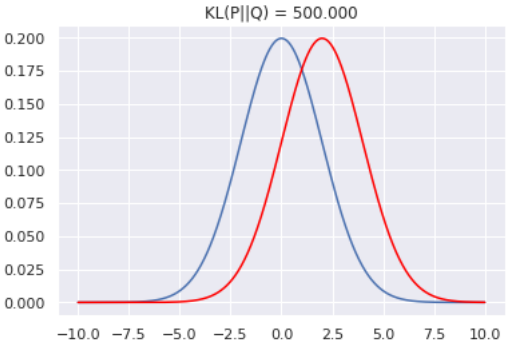
1. to construct a probability distribution over pairs of high-dimensional objects in such a way that similar objects are assigned a higher probability while dissimilar points are assigned a lower probability.



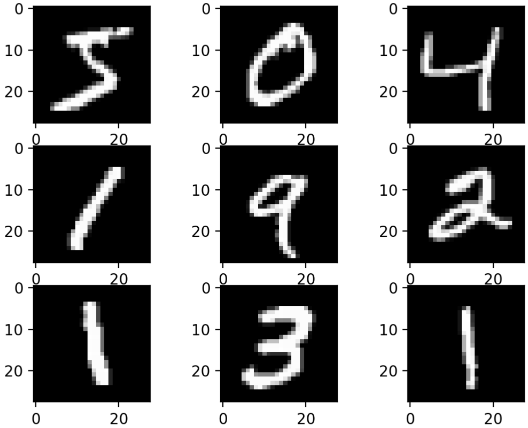
t-SNE

t-SNE algorithm (two stages) Wikipedia [2021b]

1. to define a similar probability distribution over the points in the low-dimensional map, and it minimizes the **Kullback–Leibler divergence** between the two distributions with respect to the locations of the points in the map.

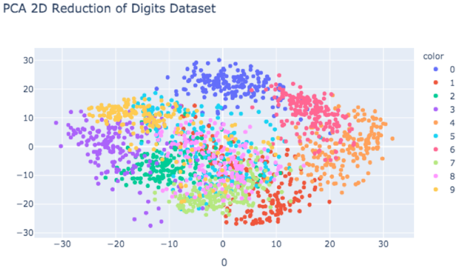


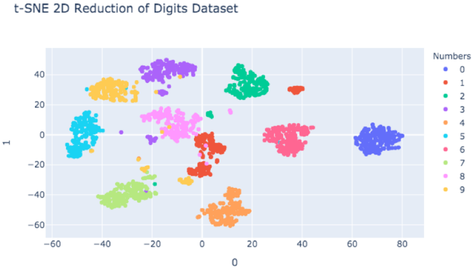
t-SNE - MNIST Digits Dataset[1]



[1] https://observablehq.com/@robert-browning/t-sne-t-distributed-stochastic-neighbor-embedding

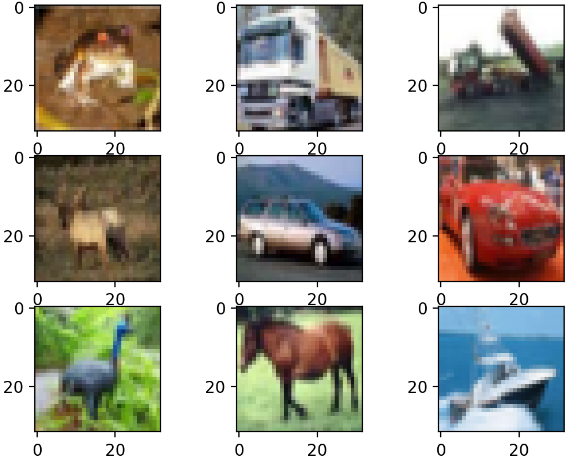
t-SNE - MNIST Digits Dataset[1]



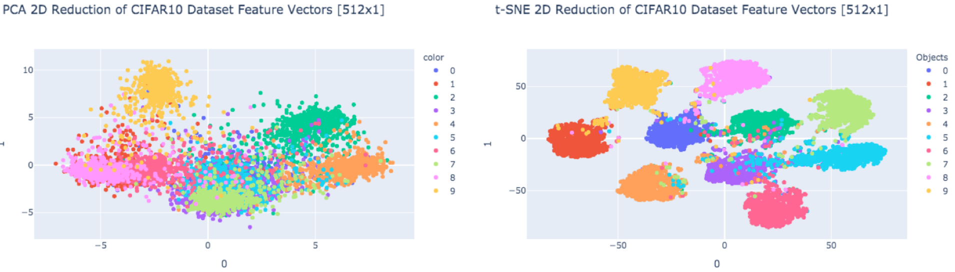


[1] `https://observablehq.com/@robert-browning/t-sne-t-distributed-stochastic-neighbor-embedding`

t-SNE - CIFAR10 Dataset



t-SNE - CIFAR10 Dataset

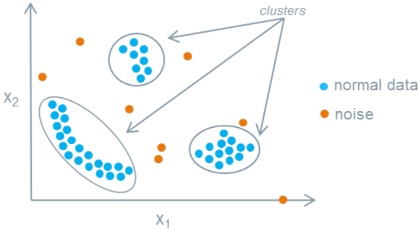


Examples of dimensionality reduction

Let us do Machine Learning
Code source - [Link]

Overview

Anomaly detection



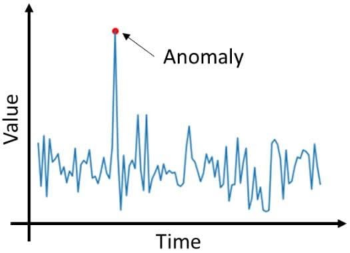


Figure: Examples of Anomaly detection

Neural network - Generative adversarial network

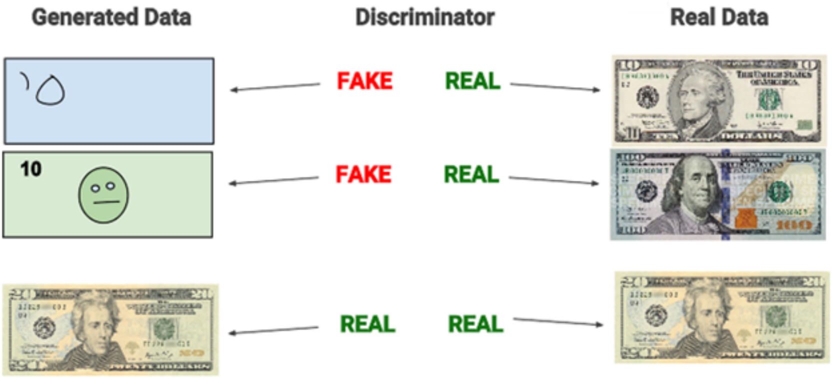


Figure: Examples of GAN

Overview

Bibliography I

Dougherty, G. (2012). *Pattern recognition and classification: an introduction*. Springer Science & Business Media.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer New York Inc.

Hinton, G. E., Sejnowski, T. J., et al. (1999). *Unsupervised learning: foundations of neural computation*. MIT press.

Kassambara, Visitor, T. t., Visitor, Agathe, Enrique, and Julie (2017). Pca - principal component analysis essentials.

Loon, A. R. v., Raut, A. S., Matthews, A. K., Sajjad, A. F., NT, A. B., Jeevan, A. M., and Guest, A. (2019). Machine learning explained: Understanding supervised, unsupervised, and reinforcement learning.

Bibliography II

Patil, P. (2018). K means clustering : Identifying f.r.i.e.n.d.s in the world of strangers.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

whuber (https://stats.stackexchange.com/users/919/whuber). Different covariance types for gaussian mixture models. Cross Validated. URL:https://stats.stackexchange.com/q/326678 (version: 2018-02-03).

Wikipedia (2021a). Dimensionality reduction.

Wikipedia (2021b). T-distributed stochastic neighbor embedding.

Thank You for Your Attention!