Unsupervised learning

Inteligencia Artificial en los Sistemas de Control Autónomo
Máster en Ciencia y Tecnología desde el Espacio

Departamento de Automática

Objectives

1. Extend unsupervised learning algorithms
2. Apply unsupervised learning to real-world problems

Bibliography

- Géron, Aurélien. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly. 2020
- Müller, Andreas C., Guido, Sarah. *Introduction to Machine Learning with Python*. O'Reilly. 2016

Table of Contents

Clustering

Set of unsupervised techniques that identify groups of data (named clusters)

- No universal definition of cluster: Centroid, medoid, dense regions, etc

Applications

- Customer segmentation, data analysis, dimensionality reduction, anomaly detection, semi-supervised learning, search engines, image segmentation

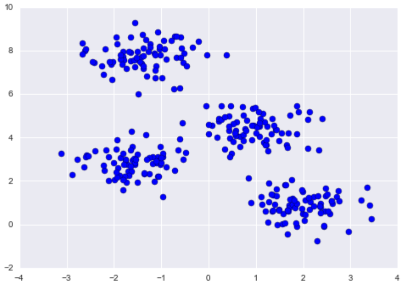
Main algorithms

- K-means, DBScan, GMM, hierarchical clustering, ...

K-means

Overview

Original data



Clustered data

(Source)

In k-means, clusters are identified by a centroid

K-means

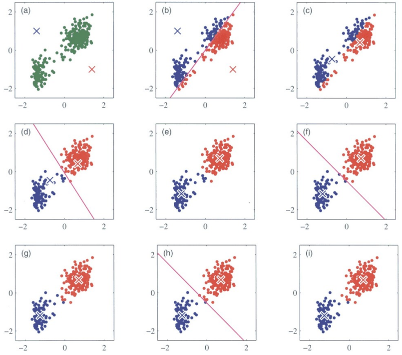
K-means algorithm (I)

K-means algorithm

1. Set k random centroids
2. Assign each data point to its closest centroid
3. Recompute centroids
4. Go to 2 until no point reassignment

k is an hyperparameter

- Number of clusters

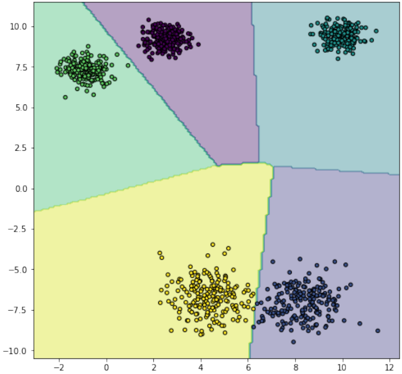


(Source)

K-means

K-means algorithm (II)

New data points are assigned to its closest centroid

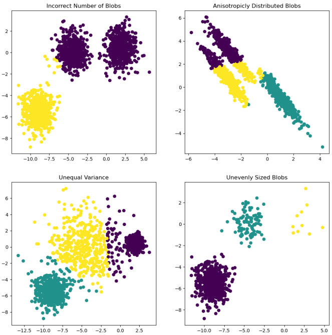


K-means

K-means limitations

K-means can fail in several conditions

- Incorrect number of clusters
- Different clusters variance
- Non-spheric clusters ⇒ normalization



(Source)

K-means

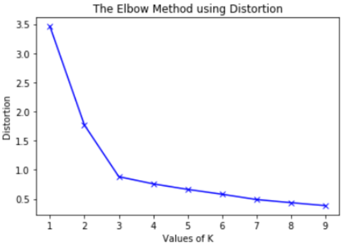
Elbow's method

Election of k

- Not a problem when domain information is available
- ... that is rarely the case

Elbow's method

1. Select $K = 1, \ldots, n$
2. Visualize performance for each k
3. Choose K where metric stabilizes

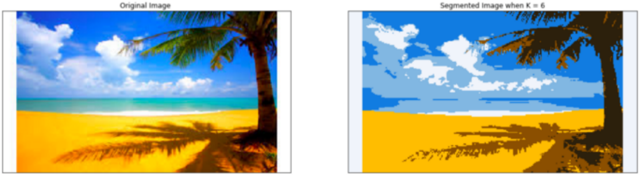


Performance measures

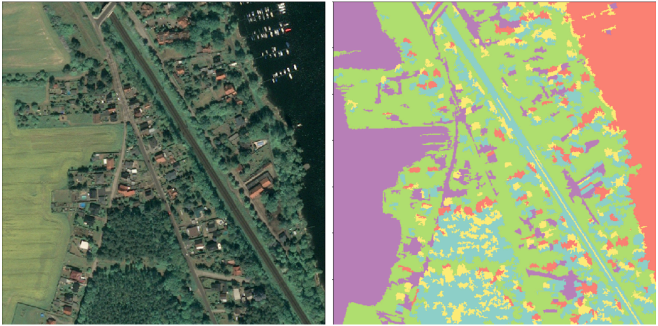
- Inertia: mean squared error between each instance and its closest centroid
- Silhouette: $(b - a)/max(a, b)$, where a mean intra-cluster distance, and b is the mean nearest-cluster distance

K-means

Application: Image segmentation



(Source)



(Source)

K-means

Application: Clustering for semi-supervised learning

Semi-supervised learning: Only a subset of the dataset is labeled

- Supervised and unsupervised learning
- Quite common in real-world applications (labels use to be expensive)

f_1	f_2	$\cdots$	f_n	y
$a_{1,1}$	$a_{2,1}$	$\cdots$	$a_{n,1}$	y_1
$a_{1,2}$	$a_{2,2}$	$\cdots$	$a_{n,2}$	
$a_{1,3}$	$a_{2,3}$	$\cdots$	$a_{n,3}$	
$a_{1,4}$	$a_{2,4}$	$\cdots$	$a_{n,4}$	y_4
$a_{1,5}$	$a_{2,5}$	$\cdots$	$a_{n,5}$	

Label propagation

1. Obtain k clusters
2. Get a representative instance of each cluster (medoid) measuring the distance to the centroid
3. Label the members of each cluster with its medoid's label

K-means

K-means: Summary

Hyperparameters	Advantages	Disadvantages
	Fast	Simple shapes
k	Few hyperparameters	Determine k
	Scalable	Random initialization

Other clustering algorithms

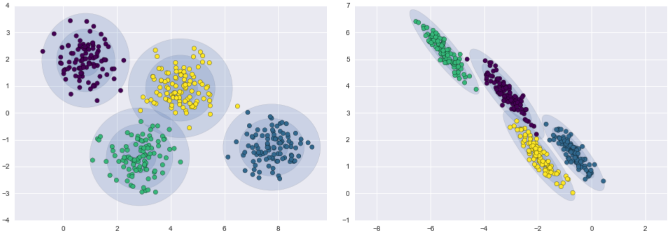
Gaussian Mixure Model (GMM) (I)

GMM is a generative clustering algorithm

- Assumes data coming from a set of multidimensional gaussian distributions

GMM fits a set $\{(\phi_i, \mu_i, \sigma_i)\}_{i=1,\ldots,k}$, where k is the number of clusters and

- ϕ is a weight
- μ is a multidimensional mean
- σ is a covariance matrix



(Source)

Other clustering algorithms

Gaussian Mixure Model (GMM) (II)

Gaussian parameters are fit with the Expectation-Maximization (E-M) algorithm

- E-M is a generalization of K-means

Expectation-Maximization algorithm

1. Init parameters randomly
2. Expectation step: Assign each instance to a cluster
   - Assignment is probabilistic
3. Maximization step: Update cluster parameters
   - Each cluster is updated using *all* the data
   - Instances contribution to a cluster parameters is weighted by the probability that it belongs to it
4. Go to 2

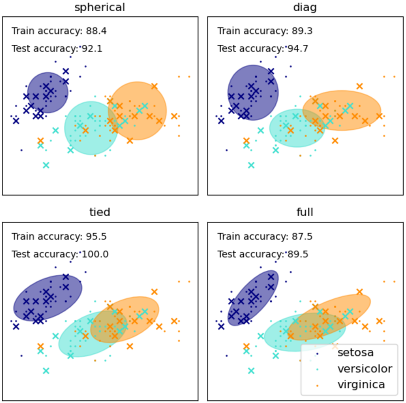
GMM can be seen as a fuzzy clustering algorithm

Other clustering algorithms

Gaussian Mixure Model (GMM) (III)

Covariance matrices can be constrained

- Full: No restriction
- Spherical: Sheprical shapes, different diameters
- Diag: Ellipsoidal shapes, axes parallel to the coordinate system
- Tied: Same ellipsoidal shapes, size and orientation

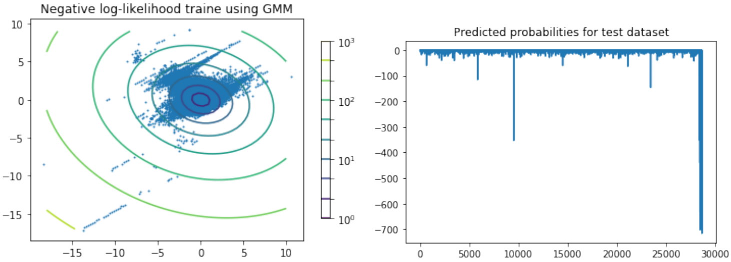


Other clustering algorithms

GMM for anomaly detection

GMM provides a probability of an instance to belong to a cluster

- This can be used to detect anomalies
- Just assign a probability threshold



(Source)

Other clustering algorithms

GMM: Summary

Hyperparameters	Advantages	Disadvantages
Number of clusters	Probabilistic clustering	Number of clusters
Covariance matrix type	Generative model	Gaussian data
	Anomaly detection	Sensitive to outliers

Other clustering algorithms

DBSCAN (I)

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

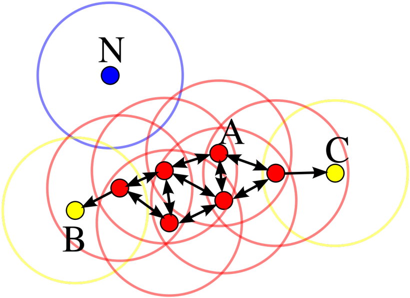
- Identifies high density regions (dense regions) in feature space
- Asumtion: Clusters form dense regions separated by empty areas

Hyperparameters

- *min_samples*: Minimun cluster size
- ϵ: Radius of a neighborhood

Type of points

- Core instance
- Frontier instance
- Outliers

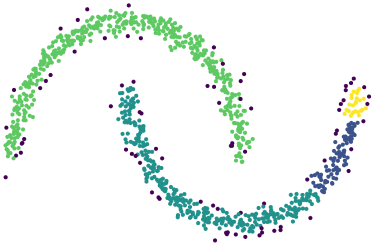


(Source)

Other clustering algorithms

DBSCAN (II)

ϵ=0.05, $min\_samples = 5$



ϵ=0.2, $min\_samples = 5$

(Source)

Other clustering algorithms

DBSCAN: Summary

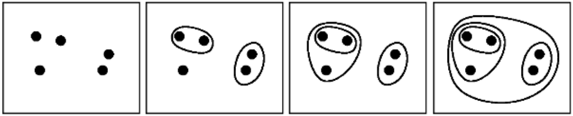
Hyperparameters	Advantages	Disadvantages
ϵ	Unknown number of clusters	Slower than K-means
min_samples	Scales relatively well	No model
	Almost deterministic	Clusters with different densities
	Robust to outliers	
	Anomaly detection	

Other clustering algorithms

Agglomerative clustering (I)

Agglomerative clustering

1. Initially, each instance forms a cluster
2. Merge the two most similar clusters according to a metric
3. Repeat 2 until a stop criterion is satisfied



We need a similarity measure between two clusters

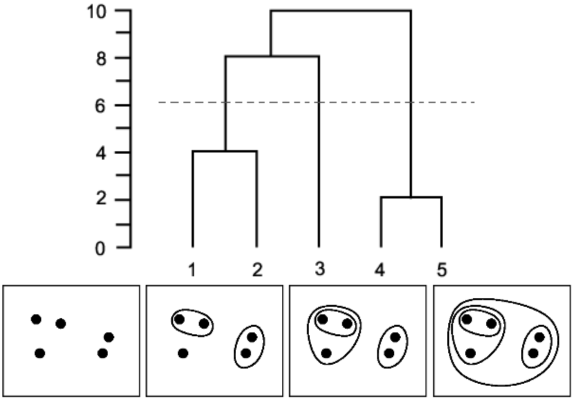
- *Ward*: Variance within merged clusters. Leads to equally sized clusters
- *Average*: Average distances
- *Complete*: Maximun distance
- *Single*: Minimum distance

Other clustering algorithms

Agglomerative clustering (II)

Agglomerative clustering is a special case of hierarchical clustering

Dendrogram

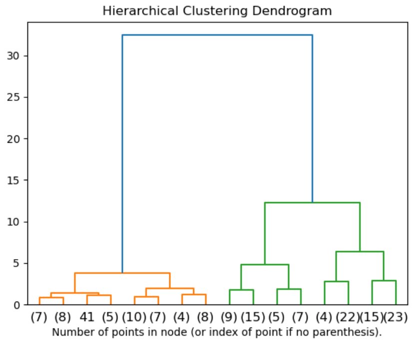


(Source)

Other clustering algorithms

Agglomerative clustering (III)

Iris dataset dendrogram



(Source)

Other clustering algorithms

Agglomerative clustering: Summary

Hyperparameters	Advantages	Disadvantages
Similarity	Complex shapes Hierarchical clustering	Slow No model

Anomaly detection

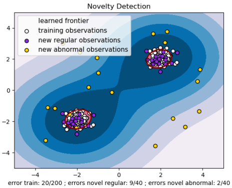
Two related concepts

- Outlayer detection and novelty detection

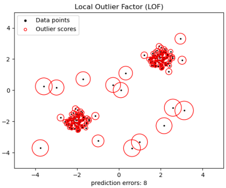
Adaptation of clustering and classification algorithms

- PCA, GMM, autoencoders, etc

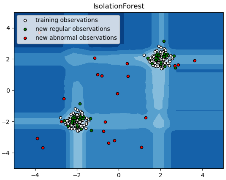
One-Class SVM



LOF



Isolation Forest



(Source)

Dimensionality reduction

Main approaches (I)

Two main approaches to dimensionality reduction: Projection and manifold learning

Projection

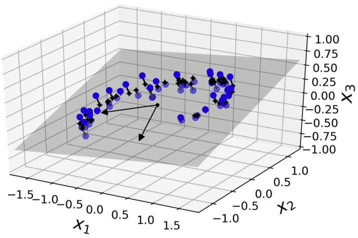


Figure 8-2. A 3D dataset lying close to a 2D subspace

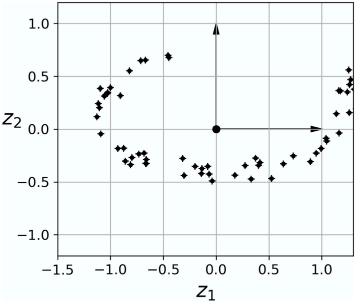


Figure 8-3. The new 2D dataset after projection

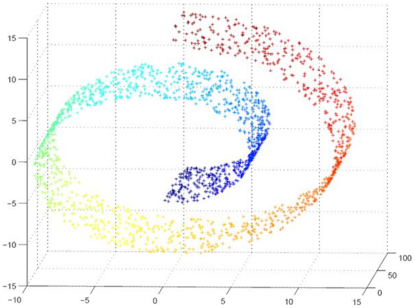
(Source)

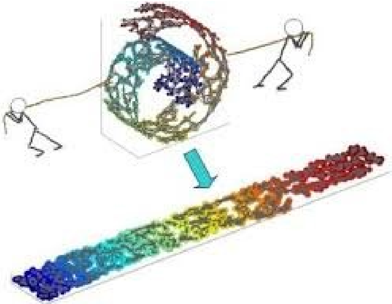
Algorithms: Principal Components Analysis (PCA), kernelized PCA, ...

Dimensionality reduction

Main approaches (II)

Manifold learning





(Source)

Manifold learning algorithms

- Isomap, T-distributed Stochastic Neighbor Embedding (t-SNE), Multi-dimensional Scaling (MDS), Locally Linear Embedding (LLE), ...

Dimensionality reduction

Principal Components Analysis

PCA create a new coordinate system

- New axes capture maximum variance and are orthogonal
  - They are named **principal components**
- The amount of variance captured by each principal component is captured
- PCA does not change the original dimensionality

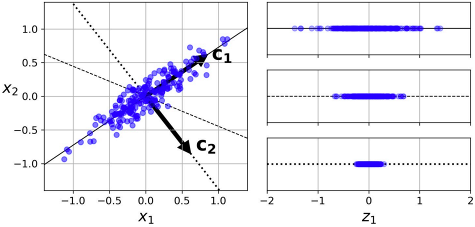
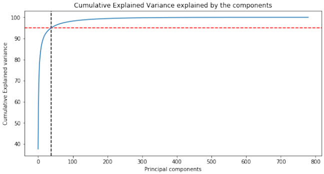


Figure 8-7. Selecting the subspace to project on

(Source)



(Source)

Dimensionality reduction

Principal Components Analysis: Applications (I)

PCA application: Image compression

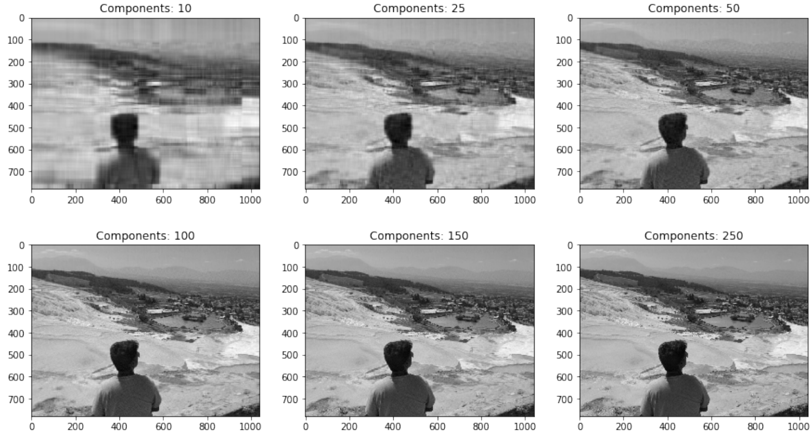
Original image

Compressed image (38 dimensions)

(Source)

Dimensionality reduction

Principal Components Analysis: Applications (II)



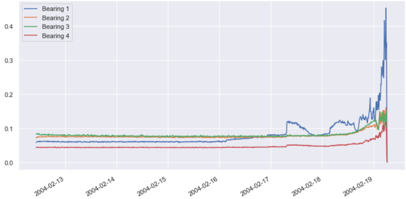
(Source)

Dimensionality reduction

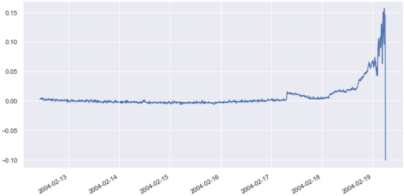
Principal Components Analysis: Applications (III)

Application: Anomaly detection to predict bearing failure, vibrations of four bearings

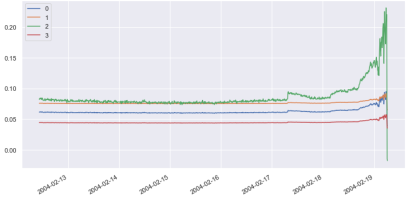
Vibration time series



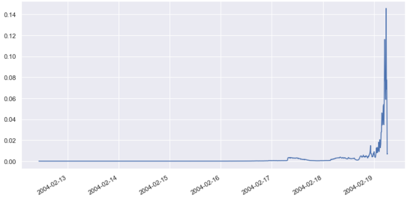
First component



Reconstructed time series



Reconstruction error



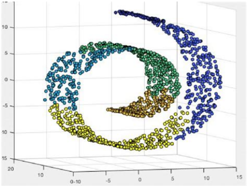
(Source)

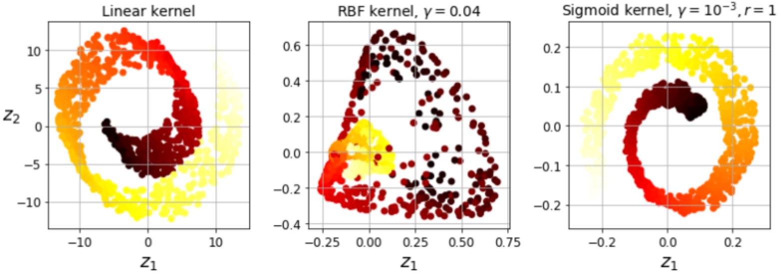
Dimensionality reduction

Kernel PCA

The kernel trick applies to PCA

- kPCA captures non-linear structures





(Source)

Dimensionality reduction

Manifold learning

Locally Linear Embedding (LLE)

- Measures how mach each training instance linearly relates to its closest neighbors preserving local relations

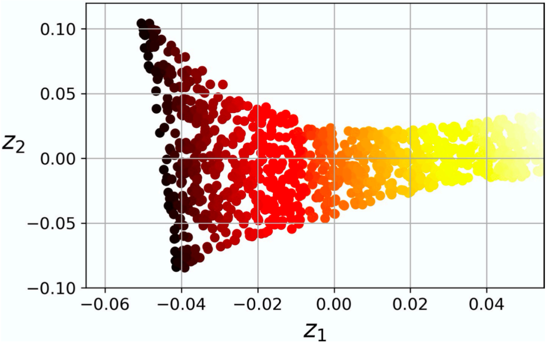


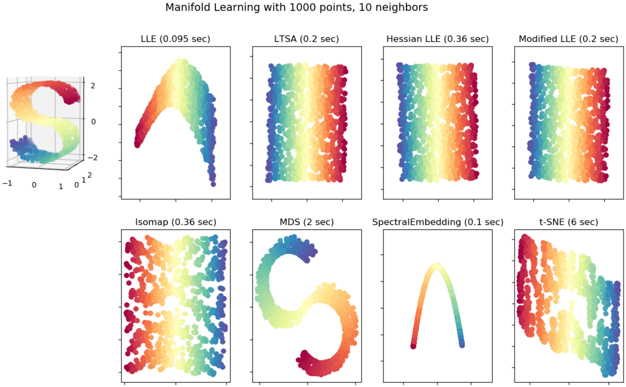
Figure 8-12. Unrolled Swiss roll using LLE

(Source)

Dimensionality reduction

Other manifold learning techniques

- Multidimensional Scaling (MDS): Preserves distances
- Isomap: Preserves geodesic distance
- t-Distributed Stochastic Neighbor Embedding (t-SNE): Preserves local distances and keep dissimilar instances apart



(Source)

Dimensionality reduction

Manifold learning demo

(Interactive demo)