

Unsupervised learning

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

Departamento de Automática

Objectives

I. TODO

Bibliography

- TODO Bishop, Christopher M. Pattern Recognition and Machine Learning. 2nd edition. Springer-Verlag. 2011
- TODO Müller, Andreas C., Guido, Sarah. Introduction to Machine Learning with Python. O'Reilly. 2016

Table of Contents

1. Clustering

- Applications

2. K-means

- Overview
- K-means algorithm
- K-means limitations
- Elbow's method
- Application: Image segmentation
- Application: for semi-supervised learning

- Summary

3. Other clustering algorithms

- GMM
- DBSCAN
- Summary
- Agglomerative clustering
- Agglomerative clustering: Summary

4. Dimensionality reduction

- PCA

Clustering

K-means, agglomerative clustering, DBSCAN and GMM

Clustering

Applications

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

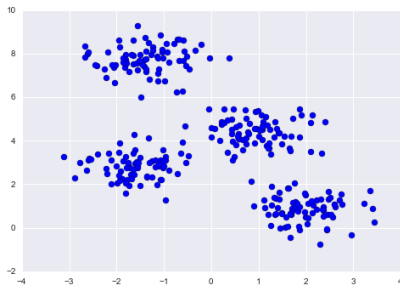
Main algorithms

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

K-means

Overview

Original data



(Source)

Clustered data



In k-means, clusters are identified by a centroid

K-means

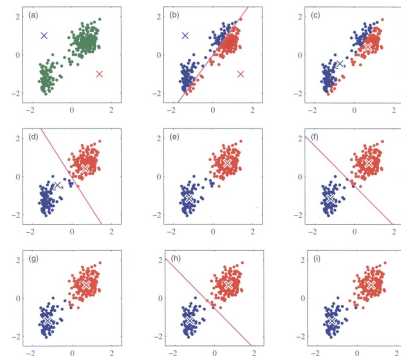
K-means algorithm (I)

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

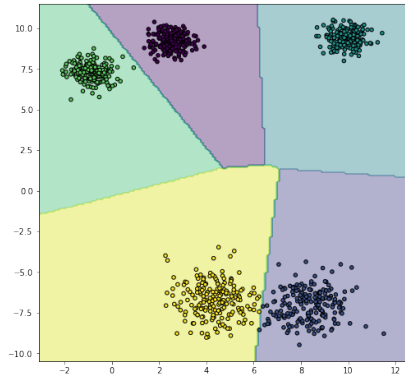
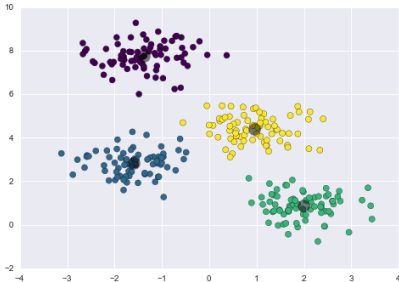


(Source)

K-means

K-means algorithm (II)

New data points are assigned to its closest centroid

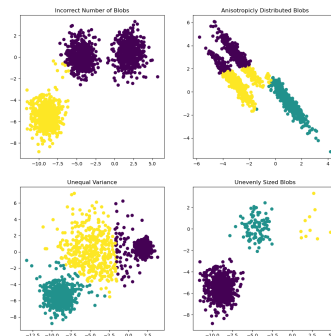


K-means

K-means limitations

K-means can fail in several conditions

- Incorrect number of clusters
- Different clusters variance
- Non-spheric clusters \Rightarrow normalization



(Source)

K-means

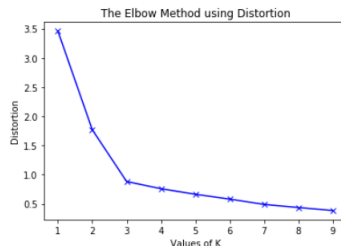
Elbow's method

Election of k

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

Elbow's method

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



Performance measures

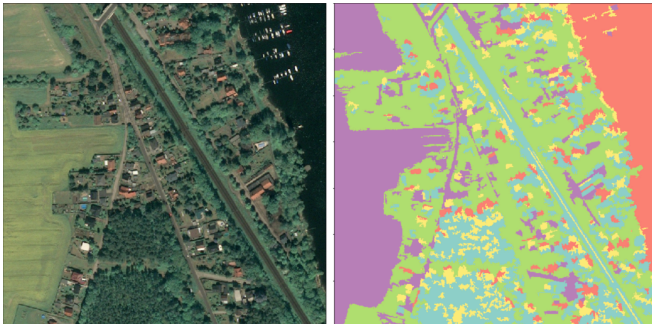
- Inertia: mean squared error between each instance and its closest centroid
- Silhouette: $(b - a) / \max(a, b)$, where a mean intra-cluster distance, and b is the mean nearest-cluster distance

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	\dots	f_n	γ
$a_{1,1}$	$a_{2,1}$	\dots	$a_{n,1}$	γ_1
$a_{1,2}$	$a_{2,2}$	\dots	$a_{n,2}$	
$a_{1,3}$	$a_{2,3}$	\dots	$a_{n,3}$	
$a_{1,4}$	$a_{2,4}$	\dots	$a_{n,4}$	γ_4
$a_{1,5}$	$a_{2,5}$	\dots	$a_{n,5}$	

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

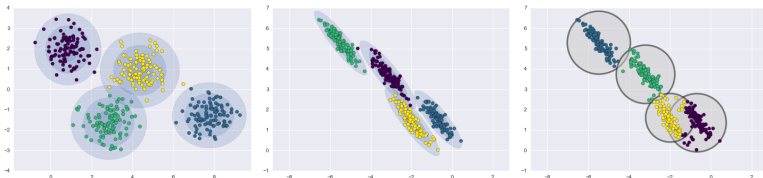
Hyperparameters	Advantages	Disadvantages
k	Fast Few hyperparameters Scalable	Simple shapes Determine k Random initialization

Other clustering algorithms

Gaussian Mixture Model (GMM)

GMM builds a probabilistic model of our data

- GMM is a generative clustering algorithm
- Assumes data coming from a set of multidimensional gaussian distributions
 - GMM fits a set $\{(\mu_i, \sigma_i)\}_{i=1, \dots, K}$
 - μ is a vector
 - σ is a covariance matrix



(Source)

Other clustering algorithms

DBSCAN (I)

DBSCAN: Density-Based Spatial Clustering of Applications with Noise

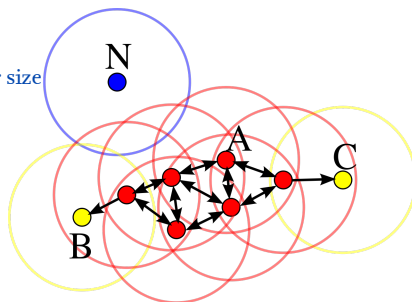
- Identifies high density regions (dense regions) in feature space
- Asumtion: Clusters form dense regions separated by empty areas

Hyperparameters

- ϵ : Radius of a neighborhood
- `min_samples`: Minumun cluster size

Type of points

- Core instance
- Outliers

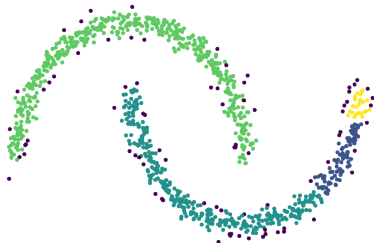


(Source)

Other clustering algorithms

DBSCAN (II)

$\epsilon=0.05$, $\text{min_samples} = 5$



$\epsilon=0.2$, $\text{min_samples} = 5$



(Source)

Other clustering algorithms

DBSCAN: Summary

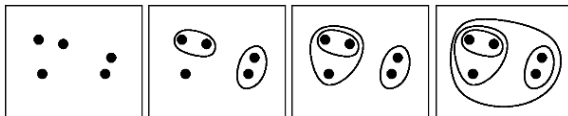
Hyperparameters	Advantages	Disadvantages
ϵ	No explicit number of clusters	Slower than K-means
min_samples	Scales relatively well	Clusters with different densities
	Almost deterministic	
	Robust to outliers	

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



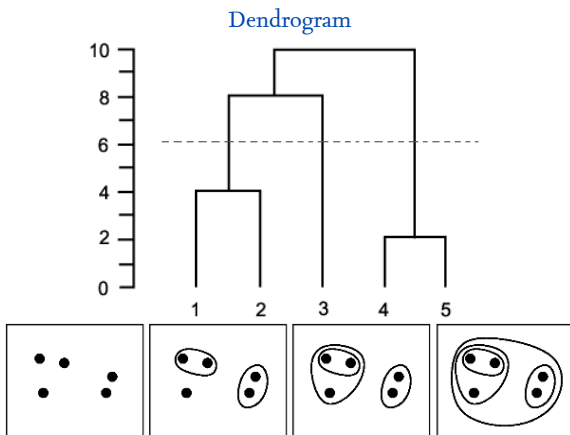
We need a similarity measure between two clusters

- Ward: Minimizes variance within merged clusters. Leads to equally sized clusters
- Average: Minimizes average distances between their points
- Complete: Minimizes maximum distance between their points

Other clustering algorithms

Agglomerative clustering (II)

Agglomerative clustering is a special case of hierarchical clustering



(Source)

Unsupervised learning

Inteligencia Artificial en los Sistemas de Control Autónomo

Algorithms

K-means: Summary

Hyperparameters	Advantages	Disadvantages
	Complex shapes	
ϵ		

Algorithms

PCA and manifold learning

Algorithms

Principal Components Analysis (I)

Dimensionality reduction transforms data into more convenient representations

- Reduce data dimensionality
- Visualize multidimensional data

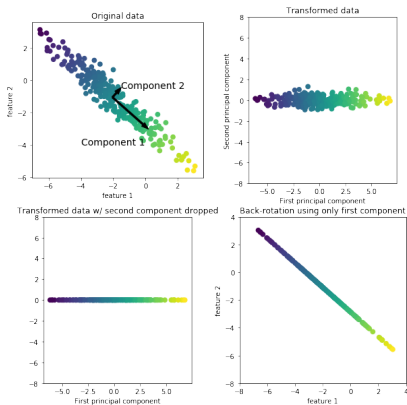
Main algorithms

- Isomap
- T-distributed Stochastic Neighbor Embedding (t-SNE)
- Principal Components Analysis (PCA)

Algorithms

Principal Components Analysis (II)

PCA maximizes data variance



(Source)

Algorithms

Principal Components Analysis (III)

Example: Hand-written digits recognition

- Images of hand-written digits
- 8x8 images (64 dimensions)
- 10 digits
- Classification problem

