

Unsupervised learning

Aprendizaje Automático para la Robótica
Máster Universitario en Ingeniería Industrial

Departamento de Automática

Objectives

1. Define Machine Learning (ML)
2. Delimit ML scope
3. Introduce the main ML tasks
4. Recognize problems as ML tasks

Bibliography

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

Table of Contents

I. Clustering

- Applications
- K-means
- K-means: Image segmentation
- K-means: Clustering for semi-supervised

learning

- K-means: Summary

- DBSCAN

- GMM

2. Dimensionality reduction

- PCA

Algorithms

K-means, DBSCAN and GMM

Applications

Clustering is a set of unsupervised techniques that identify groups of data (named clusters)

- No universal definition of cluster
 - Centroid, meroid, dense regions, etc

Applications

- Customer segmentation
- Data analysis
- Dimensionality reduction
- Anomaly detection
- Semi-supervised learning
- Search engines
- Image segmentation

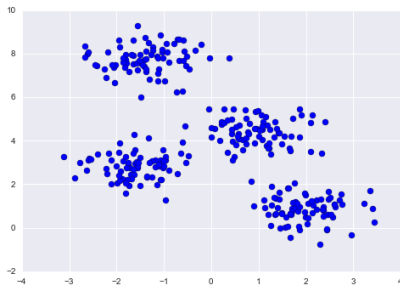
Main algorithms

- K-means, DBScan, Gaussian Mixture Models (GMM), Expectation Maximization (EM), ...

Algorithms

K-means (I)

Original data



(Source)

Clustered data



In k-means, clusters are identified by a centroid

Algorithms

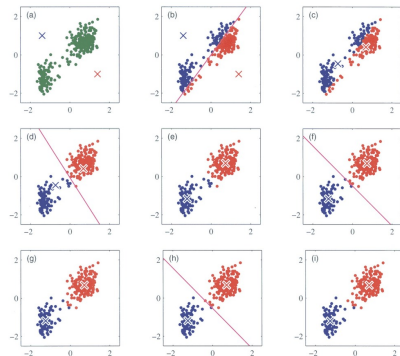
K-means (II)

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

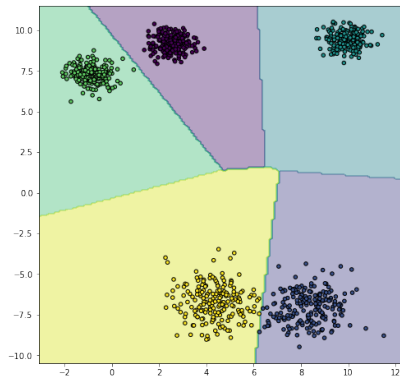


(Source)

Algorithms

K-means (III)

New data points are assigned to its closest centroid

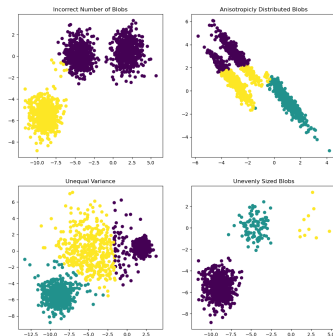


Algorithms

K-means (IV)

K-means can fail in several conditions

1. Incorrect number of clusters
2. Different clusters “diameter”
3. Non-spheric clusters



Algorithms

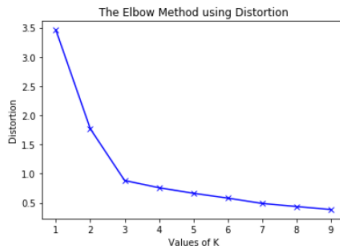
K-means (V)

K-means drawbacks

- Initial seed
- K election

Elbow 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

Algorithms

K-means: Image segmentation



(Source)



(Source)

Algorithms

K-means: 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

Algorithms

K-means: Summary

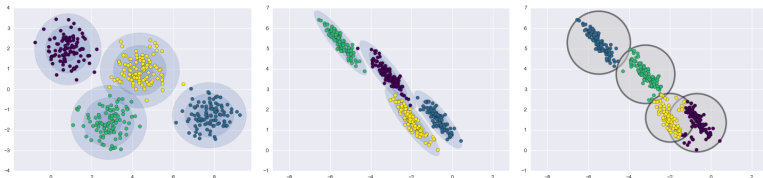
Hyperparameters	Advantages	Disadvantages
k	Fast Few hyperparameters Scalable	Simple shapes Determine k Random initialization

Algorithms

DBSCAN

GMM builds a probabilic 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



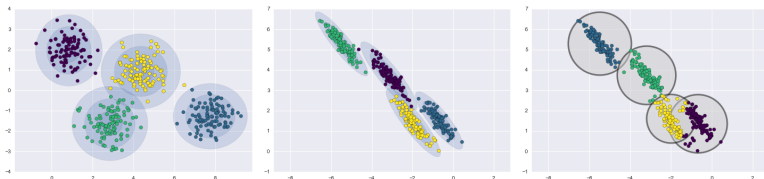
(Source)

Algorithms

Gaussian Mixture Model (GMM)

GMM builds a probabilic 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)

Algorithms

PCA and manifold learning

Algorithms

Principal Components Analysis (I)

Dimensionality reduction transforms data into more convenient representations

- Reduce data dimensionality
- Visualize multidimensional data

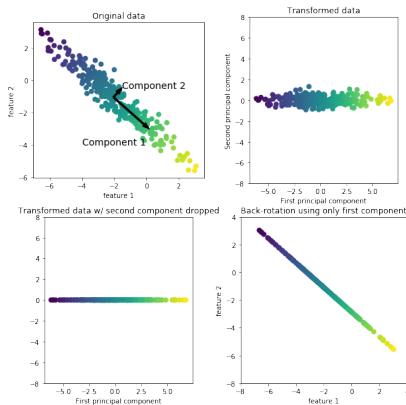
Main algorithms

- 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

