Machine Learning Lesson 1 Unsupervised learning

SILVERIO GARCÍA CORTÉS

ASSOCIATE PROF. UNIVERSITY OF OVIEDO

MAY 23, NAPLES

PREPARED FOR UNIV. FEDERICO II

Input Data Structure I: (e.g for classification, clustering, etc)

- X data matrix
- Each row of X is a sample
- Each sample is composed of values for different features.
- A sample is then composed of a set of values for the features that represent it. It is thus possible to represent each sample by a point in a feature space.
- The true classes of each sample are provided in the vector "y".
- (they must be known for training in supervised learning)

Data Matrix

Rows: samples

muestra $\mathbf{x}^m$

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \cdots & \cdots & \cdots & \cdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{bmatrix}$$

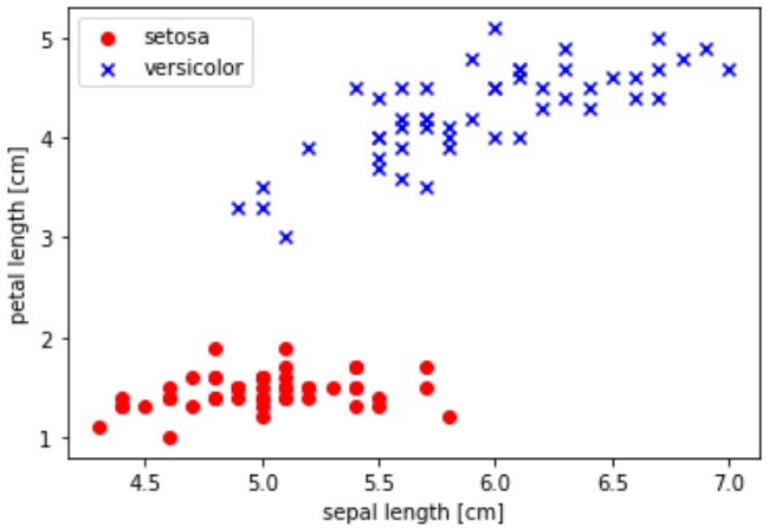
caracteristica $x_j^{(i)}$

Cols: features

True Class Vector

Supervised learning only

$$\mathbf{y} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \cdots \\ y^{(m)} \end{bmatrix}$$



Feature space: Iris dataset

Classes: 3 (50 muestras,50,50)

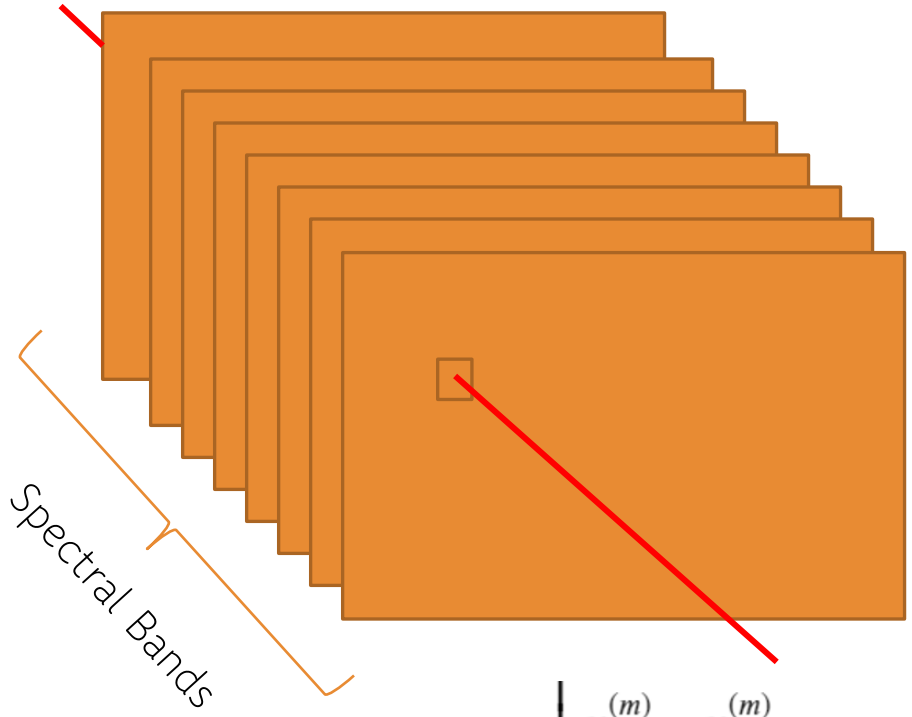
+Setosa, versicolor, virginica

Descriptors:

+Sepal Length, Sepal Width, Petal Length and Petal Width

Input data Structure II: Multispectral Imagery

Multi-band imagery can also be reshaped as the previous data matrix:



$$\begin{bmatrix} x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix}$$

Spectral vector for a pixel

$$\mathbf{X} = \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} \\ \dots & \dots & \dots & \dots \\ x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} \end{bmatrix}$$

Band 1 — Band m

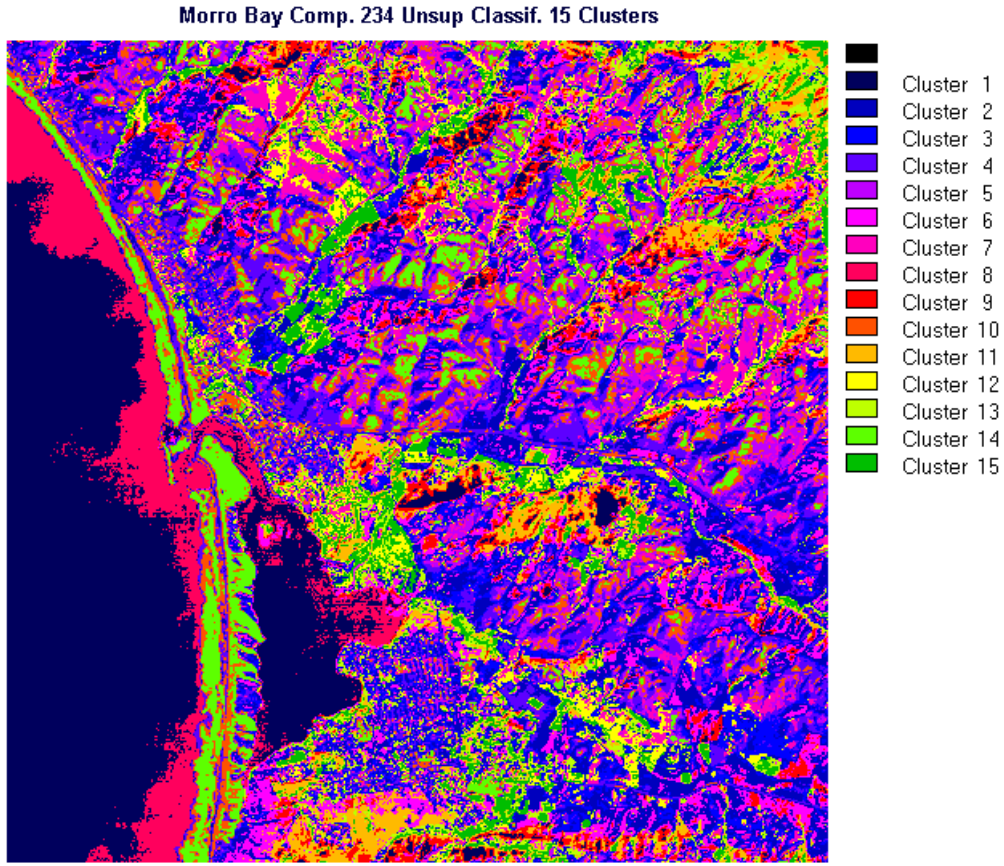
Pixel 2

Pixel n

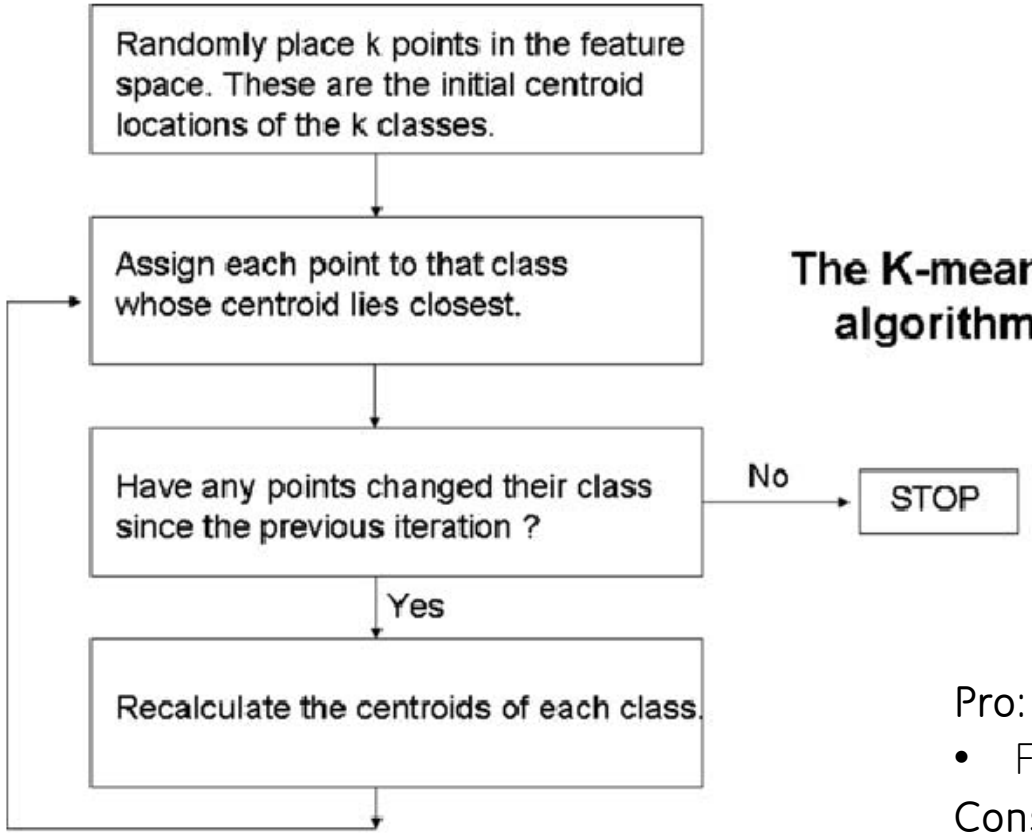
Unsupervised Learning: Clustering

Clustering

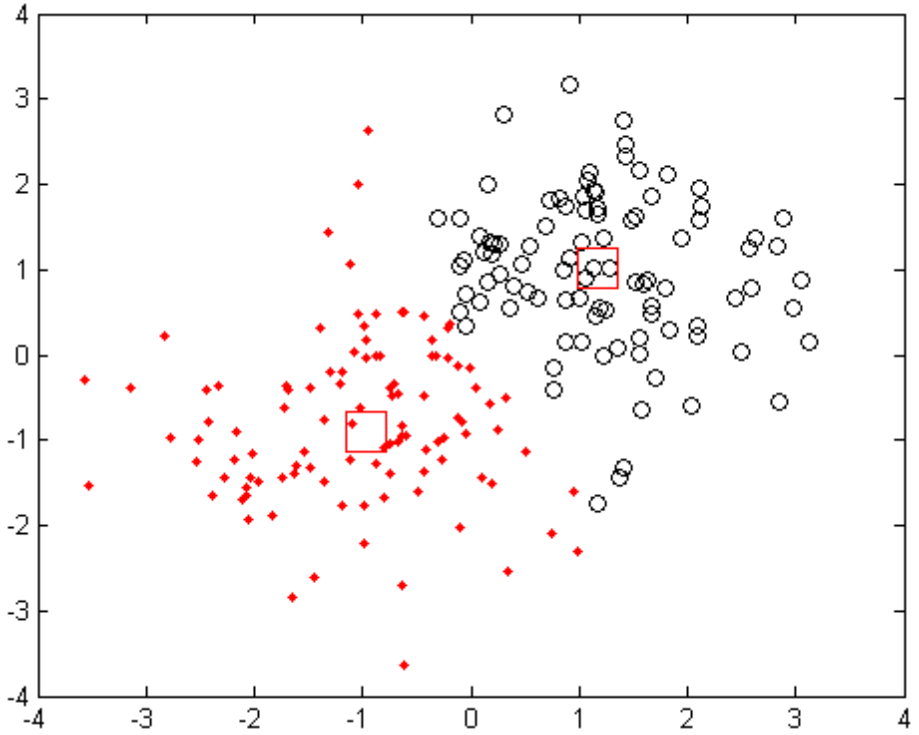
- Image clustering is a technique to group an image into clusters (units: groups of pixels) that are homogeneous with respect to one or more characteristics.
- It is applied when the features which define the interest classes are unkown or not well defined.
- In other classes this methods are used to discover groups of data with similar patterns.
- Commonly the algorithms use to start with a tentative number of clusters and each pixel is initially assigned to one of them
- During processing each sample can be reasigned to other clusters depending on some criteria. These criteria use to be given as thresholds, dispersión measurements for each cluster and some other parameters connected with a sort of "total energy" for all the samples grouping.
- In some algorithms the number of classes are fixed during all the processing and this number is set by the user.
- In other algorithms the user set an initial number of desired clusters and additional criteria that allow this number to evolve during the iterations.
- Some criteria for number of cluster evolution:
  - Número máximo de clusters. Maximum number of clusters allowed
  - Minimum Cluster center distance for agregattion.
  - Maximum cluster radius for cluster division (Splitting)
  - Minimum number of elements on a cluster (Cluster elimination)
- Different measurements can be also used to measure the "compacity" of a cluster (dispersion around the center). Standard deviation for each spectral band.
- There also exists different implementations of the same basic algorithms with different. Criteria and variants.
- Common algorithms for unsupervised learning in Satellite Remote Sensing are:
  - k-means , Isodata



CLUSTERING: K-MEANS (MOBILE MEANS ALG.)



The K-means algorithm



Global Function: $J = \sum_{j=1}^{k} \sum_{\forall i \in clase j} |x_i^j - m_j|^2$

Pro:

- Fully automatic algorithm

Cons:

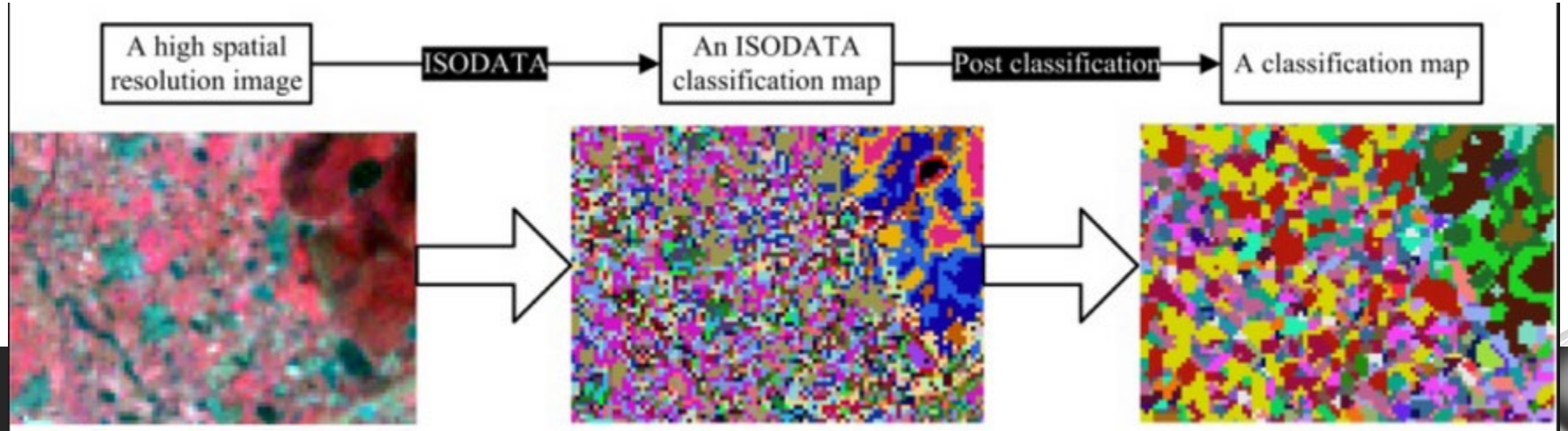
- Different posible solutions depending on initial values
- Fixed number of clusters must set a priori (before processing)

CLUSTERING: ISODATA. (ITERATIVE SELF ORGANIZING DATA ANALYSIS TECHNIQUE)

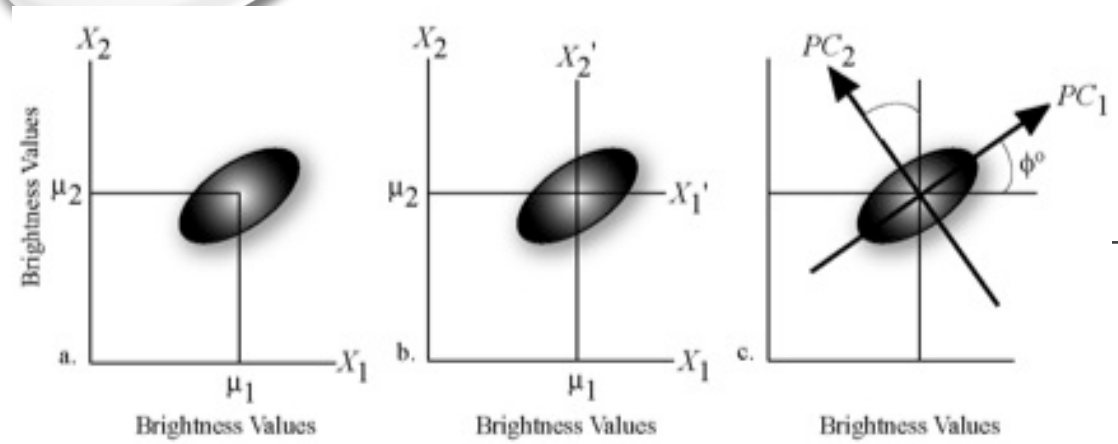
ISODATA

- THIS ALG IS MODIFICATION OF K-MEANS
- IT IS AN ITERATIVE ALGORITMO
- ARBITRARY INITIAL CLUSTER CENTER PLACEMENT (N-DIMENSIONAL VECTOR)
- NUMBER OF CLUSTERS CAN EVOLVE THROUGH ITERATIONS:
  - CLUSTER AGREGATTION IF BETWEEN CLUSTERS MEAN DISTANCE < PREDEFINIDED THRESHOLD
  - CLUSTER SPLITTING IF STÁNDARD DEVIATION > PREDEFINIDED THRESHOLD
  - CLUSTER DELETING IF MEMBER NUMBER < PREDEFINIDED THRESHOLD

- IN REMOTE SENSING (RS) APPS POSTCLASIFICACION IS COMMONLY NEEDED AFTER CLUSTERING, (MODE OR MAJORITY FILTERS)
- IN RS SOMETIMES THIS METHOD CAN SERVE TO GUIDE A FOLLOWING SUPERVISED CLASSIFICATION PROCESSING



MACHINE LEARNING: PRINCIPAL COMPONENT ANALYSIS (PCA)



$$X = P - M$$
$$P: Bandas\ espectrales$$
$$M: Matriz\ de\ medias\ de\ la\ bandas$$

$$Y = D^T \cdot X$$
$$D: Matriz\ de\ vectores\ propios$$
$$Y: Matriz\ de\ componentes\ principales$$

PCA is a mathematical technique that transforms the original bands (usually highly correlated with each other) into a series of new uncorrelated bands (linear combination of the originals) that also have decreasing variance (information) contents..

Each new band contains less variance than the previous one and so on. Commonly the first two bands contain most of the variance of the original set of bands (even more than 90%). It is possible in many cases to replace the original set of bands by the first three orthogonal components of the transformation only.

This technique is used to reduce the dimensionality of the data (especially in hyperspectral imaging), to highlight classification classes or to perform the supervised classification itself.

The eigenvector matrix is calculated from the diagonal eigenvalue matrix of the covariance matrix of the original bands. This diagonal matrix contains the eigenvalues sorted from largest to smallest eigenvalue.