Machine 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

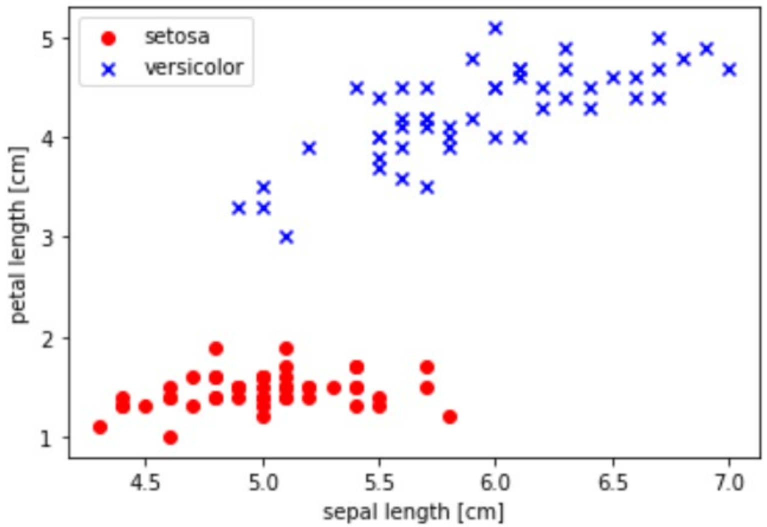
Rows: samples

muestra $\mathbf{x}^m$

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

caracteristica $x_j^{(i)}$

Cols: features



Feature space: Iris dataset

Classes: 3 (50 muestras,50,50)

+Setosa, versicolor, virginica

Descriptors:

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

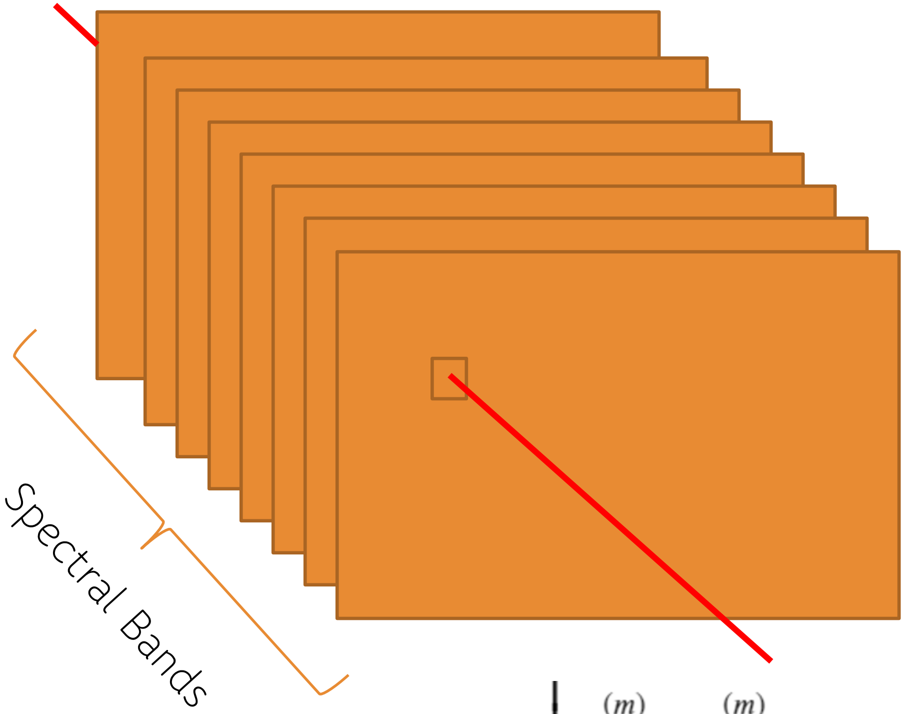
True Class Vector

Supervised learning only

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

Input data Structure II: Multispectral Imagery

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



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

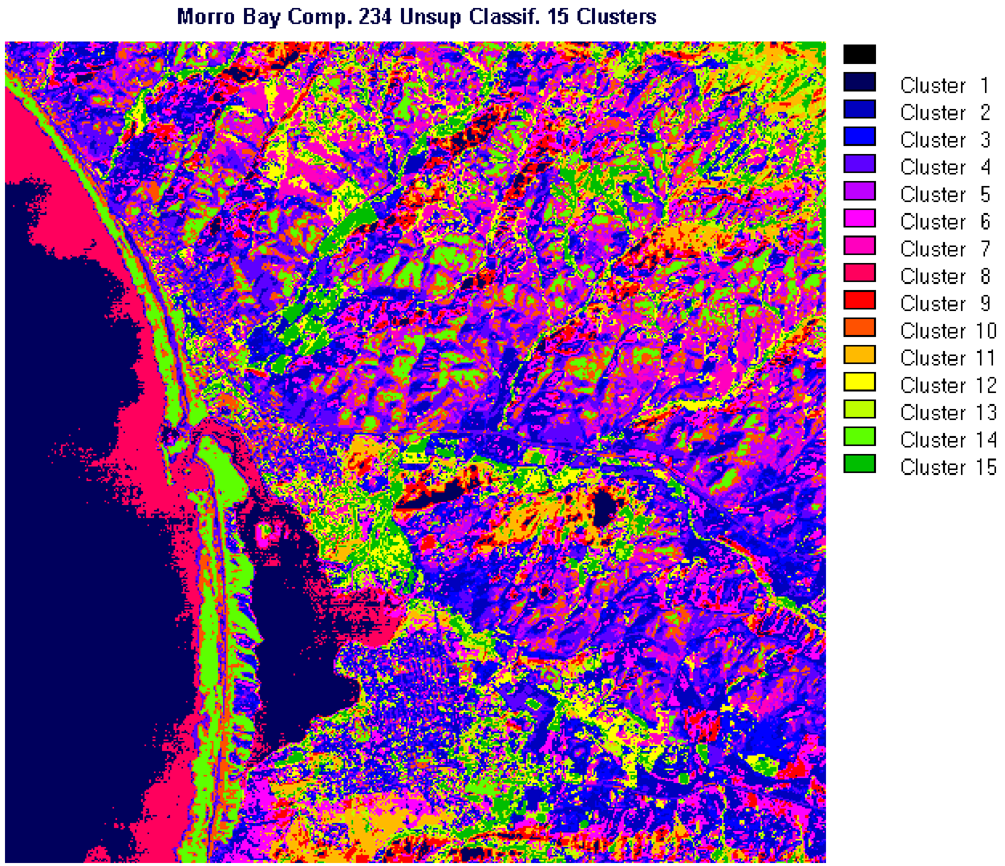
Spectral vector for a pixel

$$\mathbf{X} = \begin{array}{ccccc} \text{Band 1} & & & \text{Band m} & \\ x_1^{(1)} & x_2^{(1)} & \dots & x_n^{(1)} & \\ x_1^{(2)} & x_2^{(2)} & \dots & x_n^{(2)} & \text{Pixel 2} \\ \dots & \dots & \dots & \dots & \\ x_1^{(m)} & x_2^{(m)} & \dots & x_n^{(m)} & \text{Pixel n} \end{array}$$

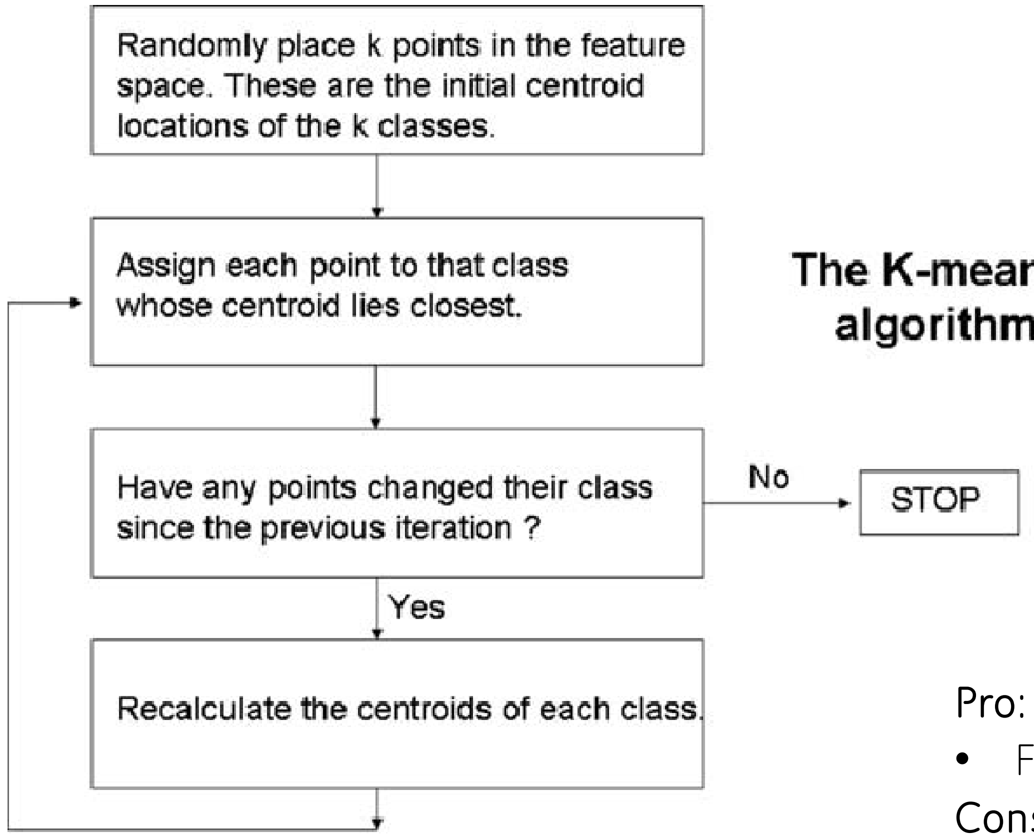
Unsupervised Learning: Clustering

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 depdending on some criterio. These criterio 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. umbrales decisión para agrupamiento, separación y eliminación de clusters
- 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 implmentations of the same basic algorithms with different. Criteria and variants.
- Common algorithms for unsupervised learning in Satellite Remote Sensing are:
  - k-means , Isodata


Morro Bay Comp. 234 Unsup Classif. 15 Clusters

CLUSTERING: K-MEANS (MOBILE MEANS ALG.)



Randomly place k points in the feature space. These are the initial centroid locations of the k classes.

Assign each point to that class whose centroid lies closest.

Have any points changed their class since the previous iteration ?

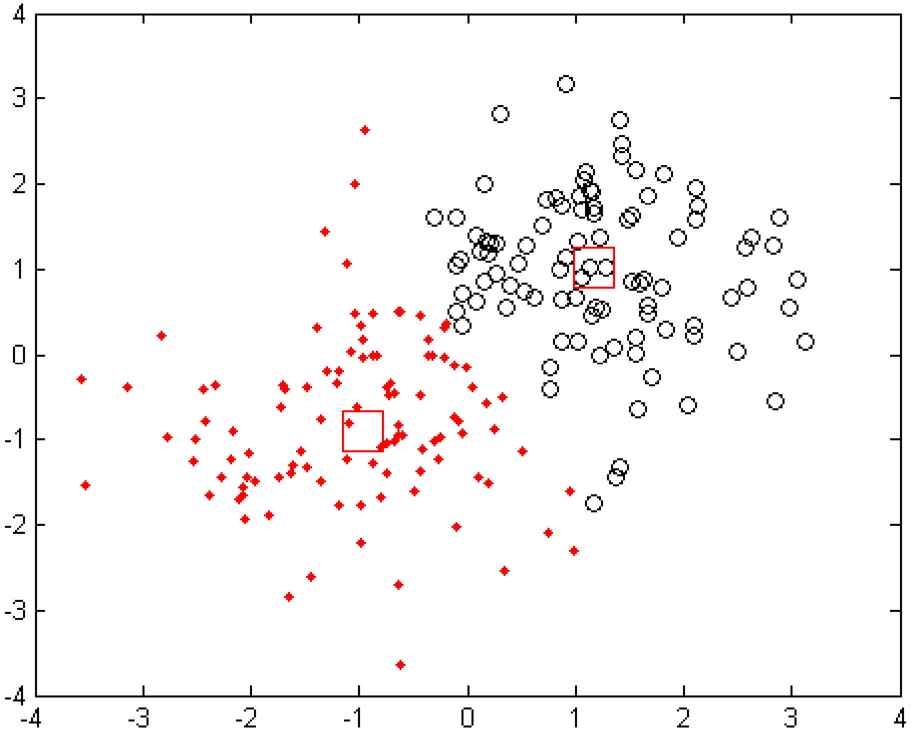
No

STOP

Yes

Recalculate the centroids of each class.

The K-means algorithm



Pro:

- Fully automatic algorithm

Cons:

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

Global Function: $J = \sum_{j=1}^{k} \sum_{\forall i \in clasej} |\mathbf{x}_i^j - \mathbf{m}_j|^2$