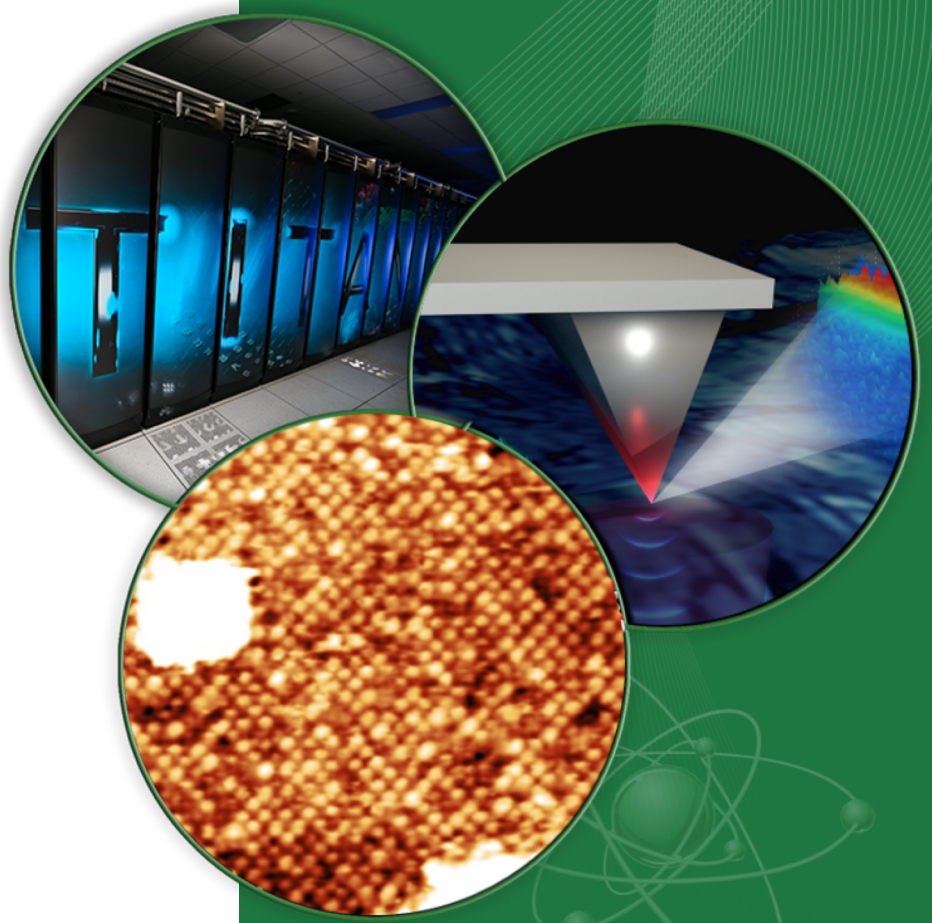
Supervised and Unsupervised Classification Methods

Rama Vasudevan

Workshop on Informatics in Advanced Measurements

Tsukuba, Japan
19th January 2017



ORNL is managed by UT-Battelle
for the US Department of Energy

OAK RIDGE National Laboratory

Outline

1. The importance of pre-processing
2. Supervised Classification
   - Support Vector Machines
   - Decision Trees
3. Unsupervised Classification
   - K-means algorithm
   - Other unsupervised methods

Pre-processing

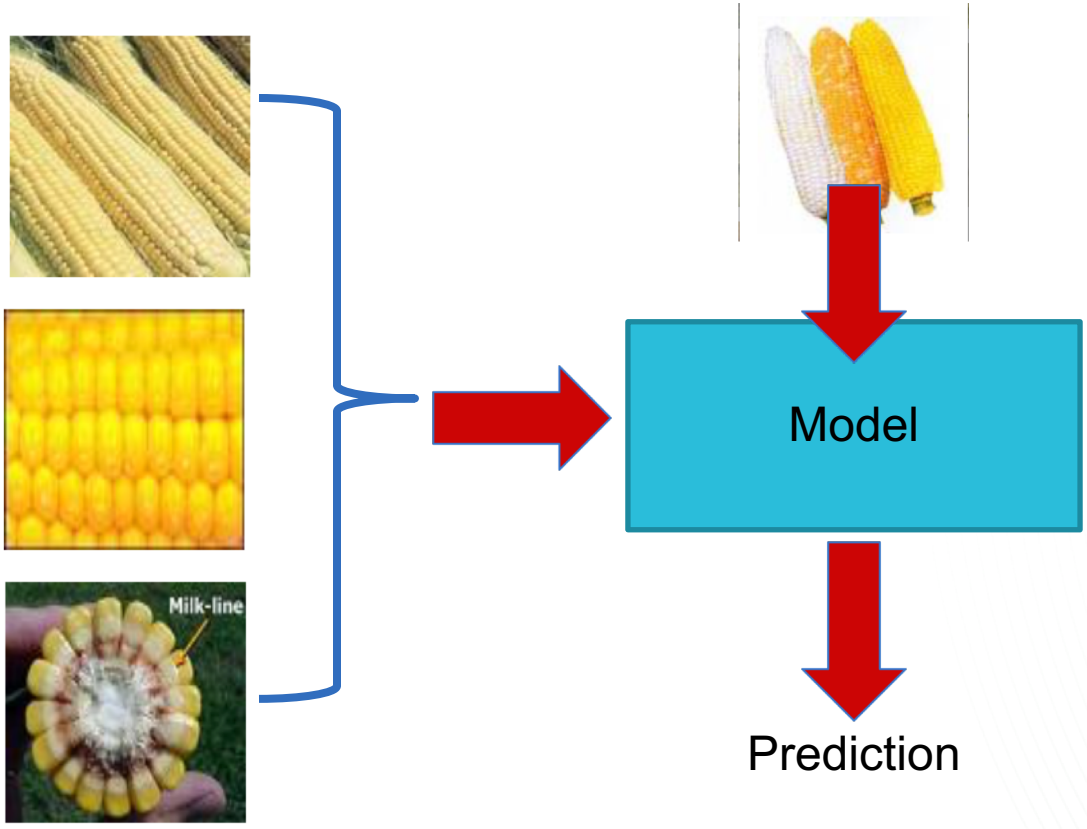
- Machine learning learns from data, without respect to what is an artifact and what is a real signal.

- As a result, care and appropriate pre-processing is required. Examples include step-changes in response, shifts of the signal in time, changes to intensity from artefacts (e.g. topographic), etc.

- Old adage: garbage in, garbage out.

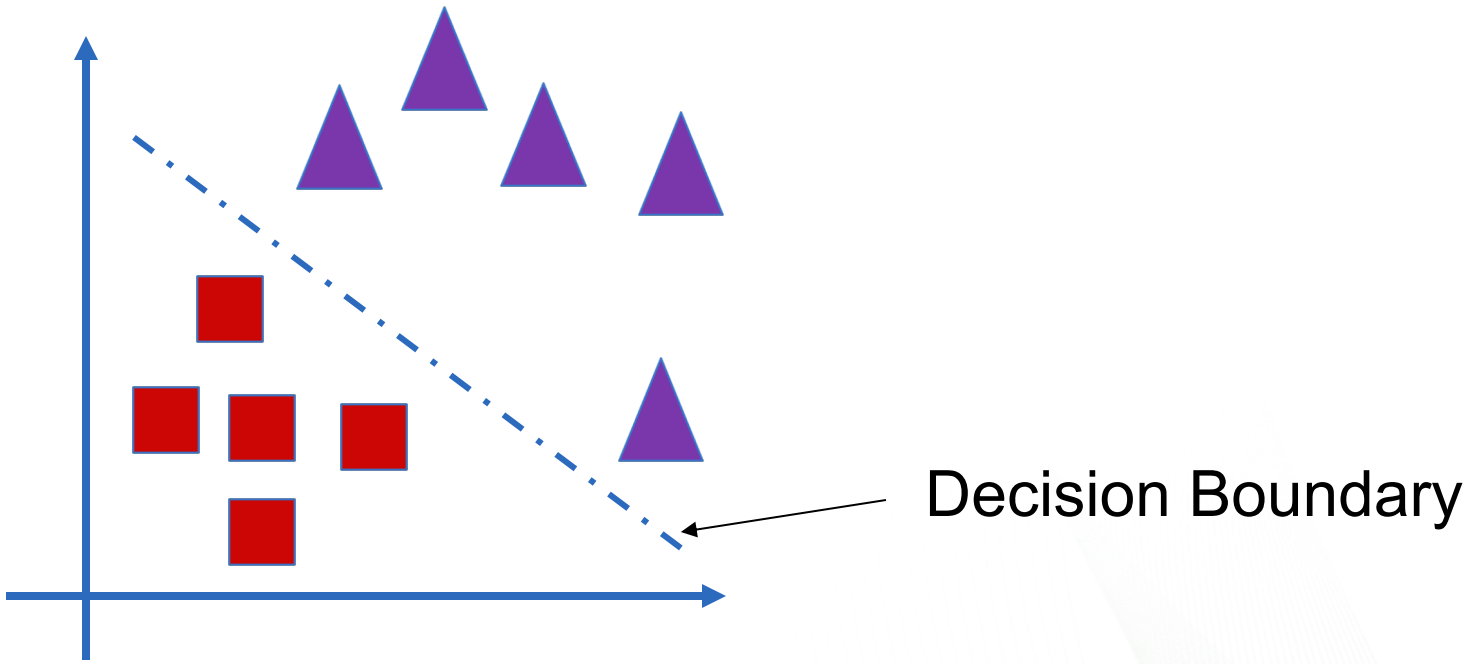
Supervised Classification

- Supervised classifiers can automatically classify data once the model has been trained on a ‘training set’.



Supervised Classification

- Support Vector Machines are one common method, and is conceptually easy to understand.

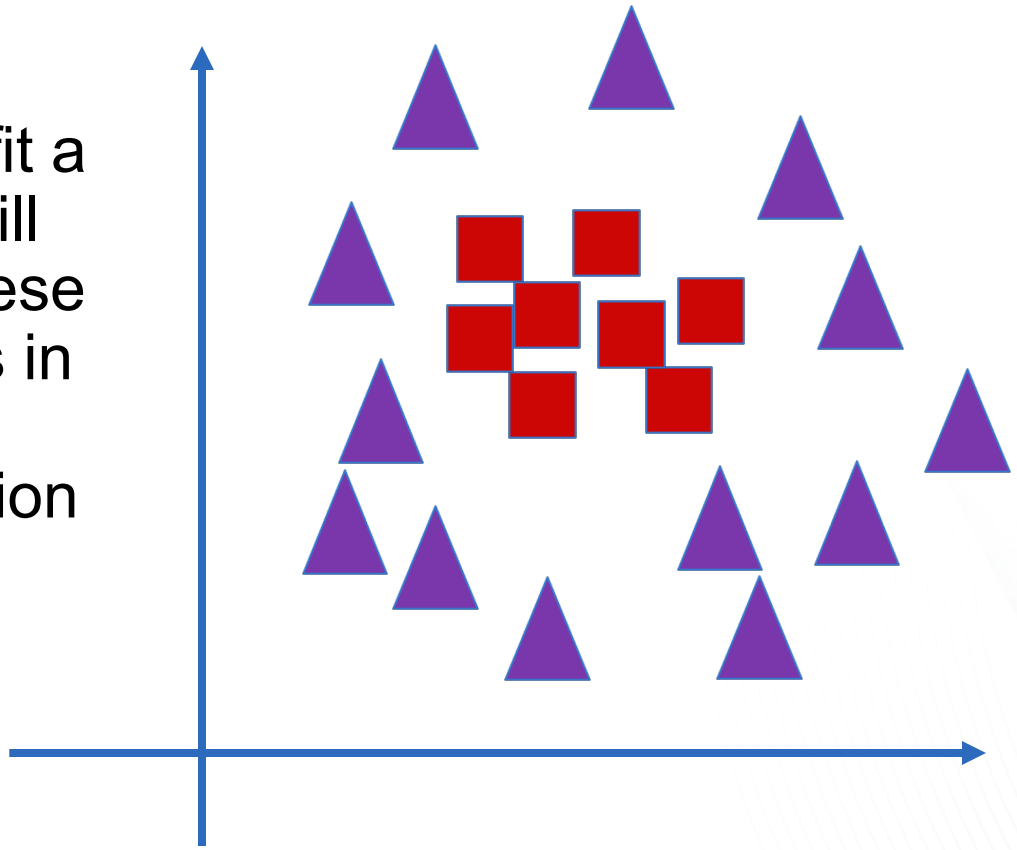


SVM are linear classifiers that fits lines (planes/hyperplanes in higher dimensions), such that the margin (separation between the line and the nearest training point) is maximized.

Supervised Classification

- But what to do when it cannot be linearly separated? E.g.,

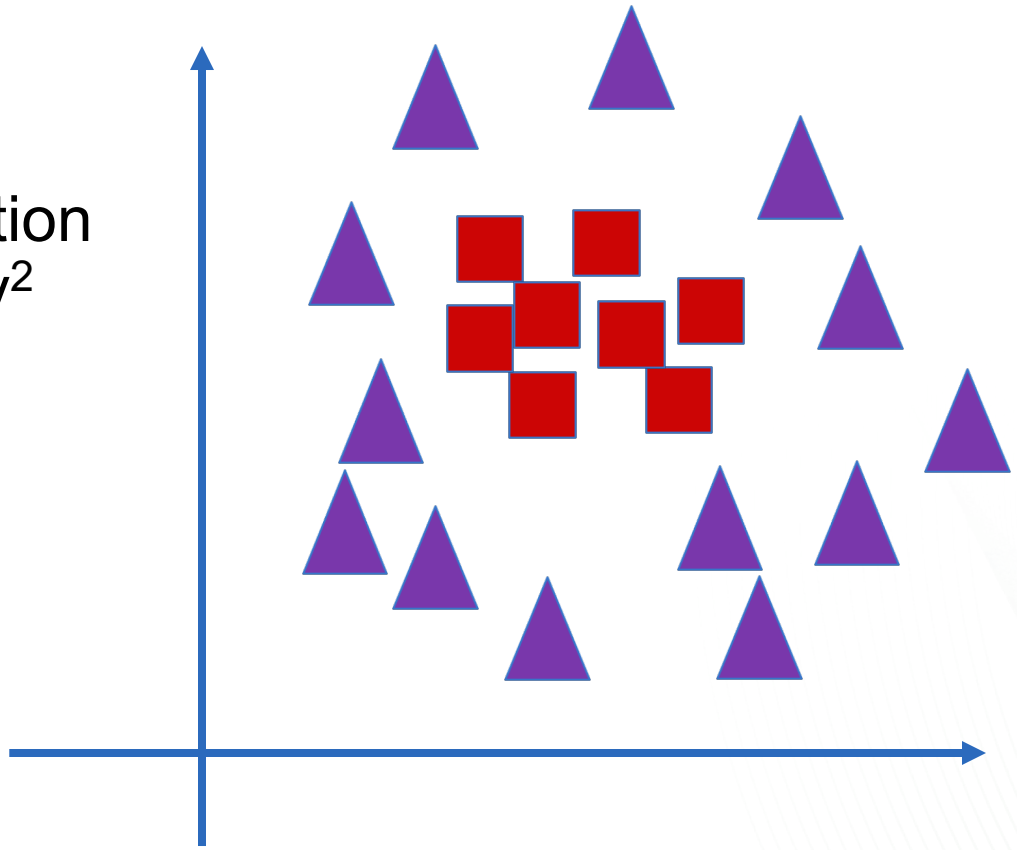
We cannot fit a line that will separate these two classes in this representation



Supervised Classification

- Solution: Transform to a space where they can be separated.

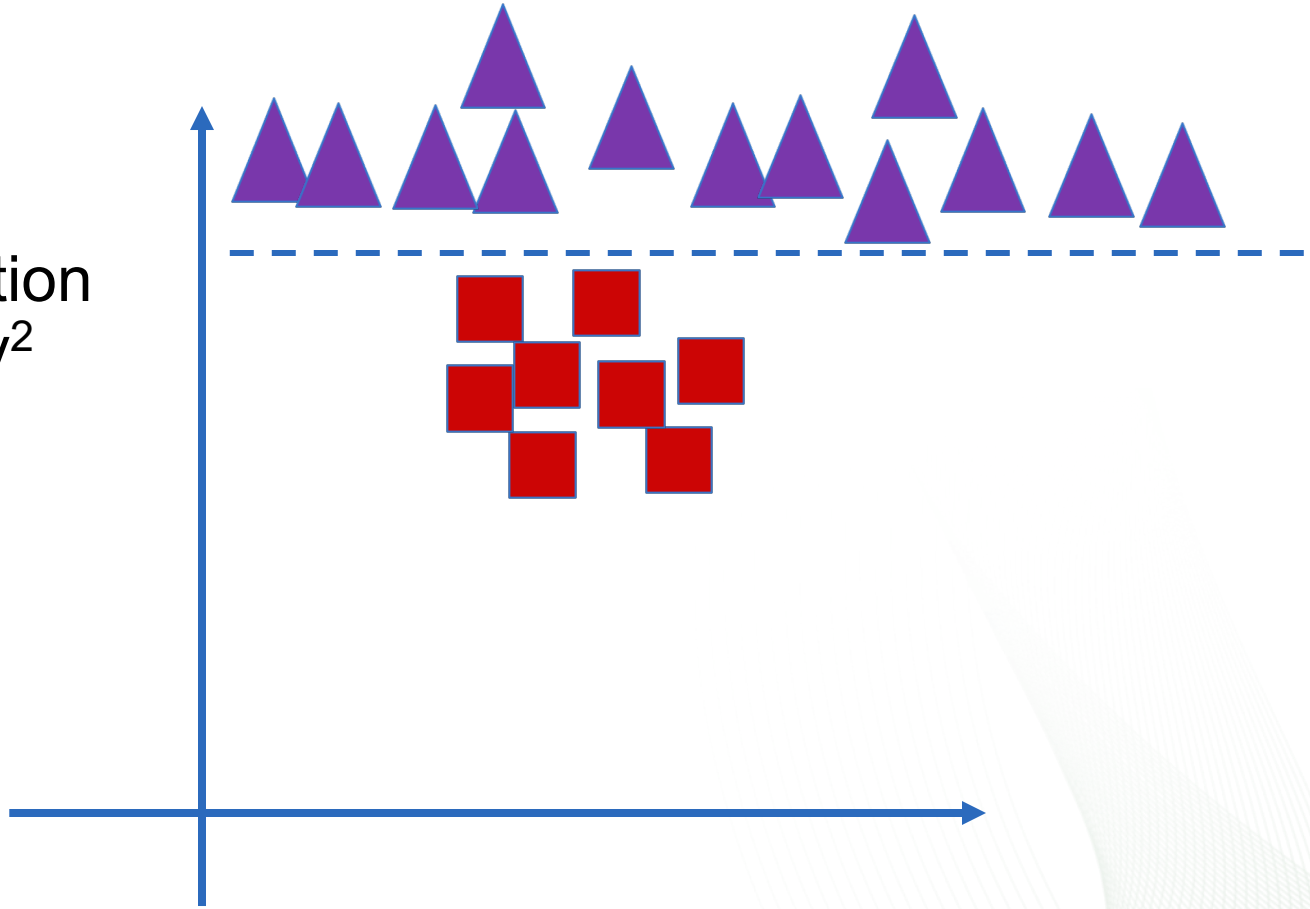
Apply a transformation
$Z = x^2 + y^2$



Supervised Classification

- Solution: Transform to a space where they can be separated.

Apply a transformation
$Z = x^2 + y^2$



Supervised Classification

- This is called the kernel trick, where a kernel function is applied to enable the separation to occur in the transformed space.

- Let's see a notebook of the use of the SVM method to classify spectral data

Decision Trees

- A very common method that is used to classify data is the decision tree
- The decision tree is easily interpretable, and essentially learns to split data based on values of features.
- Advantages: will ignore features that are unnecessary to the classification, when small they are easy to interpret, and have often been shown to be extremely good (esp. with random forests, bagging, etc.)

Decision Trees

- The decision tree is learned by using the concept of information entropy. Each split is chosen to maximize the purity of each daughter node of the tree.
- Numerous types of metrics can be used here, including KL divergence, or Gini Impurity, or Variance Reduction
- The main issue with decision trees is the tendency to overfit to the training data. Ensemble methods can reduce this tendency.

Supervised Classification

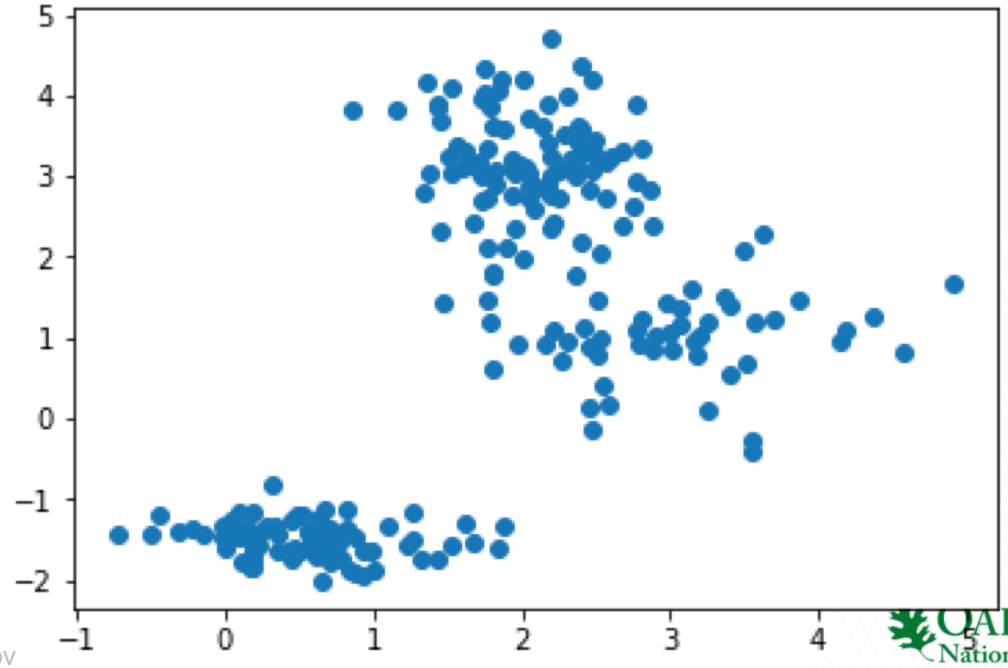
- Let's now use the notebook to use a decision tree classifier

Unsupervised Classification

- In many cases, we do not have labeled examples. In this case, we can turn towards unsupervised methods
- The most common method is k-means clustering, but there are numerous others
- In the interests of time, we will simply explore some of the other clustering methods in a notebook

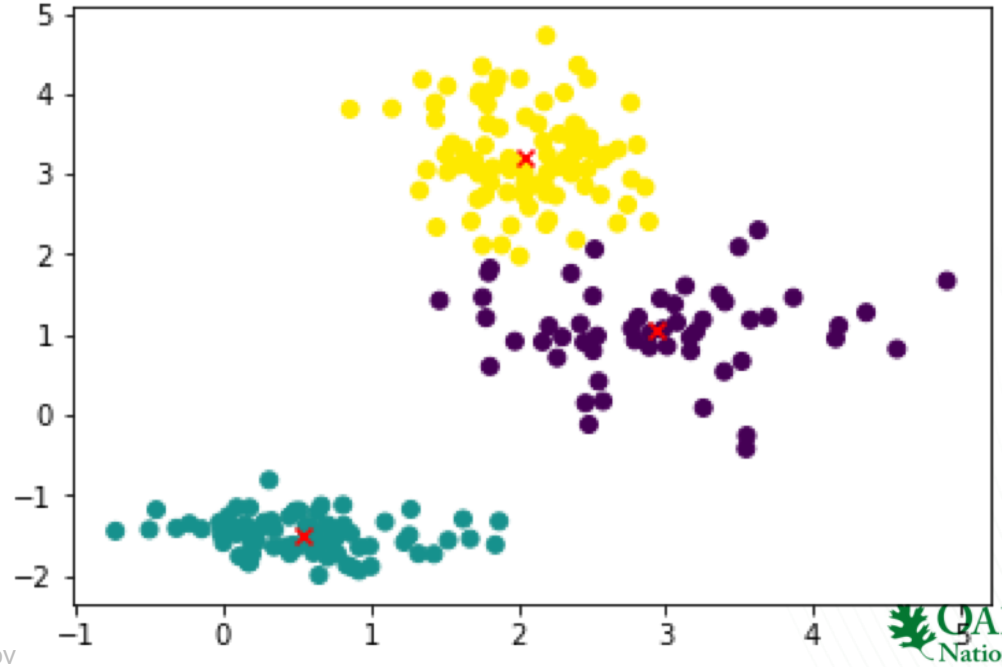
K-Means Clustering

- SVD, NMF and ICA are decompositions. But sometimes, we don't want to decompose our signal, but just group them into 'alike' sets.
- This is termed 'clustering'. The easiest and most widely used method is the k-means algorithm



K-Means Clustering

- SVD, NMF and ICA are decompositions. But sometimes, we don't want to decompose our signal, but just group them into 'alike' sets.
- This is termed 'clustering'. The easiest and most widely used method is the k-means algorithm



K-Means Clustering

- SVD, NMF and ICA are decompositions. But sometimes, we don't want to decompose our signal, but just group them into 'alike' sets.
- This is termed 'clustering'. The easiest and most widely used method is the k-means algorithm

K-means Clustering algorithm, to separate data (x_1,x_2,…,x_n) into *k* clusters

$$\arg\min_{\boldsymbol{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}\in S_i} \left|\left|\mathbf{x} - \boldsymbol{\mu}_i\right|\right|^2$$ where μ_i is the mean of points in S_i

(Determine **S** = {S_1,S_2,…S_k}, such that within cluster sum of squares is minimized)

Unsupervised Classification

- Let's now use the notebook to do clustering on the available datasets.