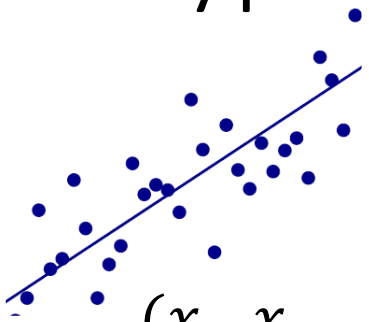
Supervised vs Unsupervised Learning

By Francisco Mendoza

mentofran@gmail.com

Type of problems, data types

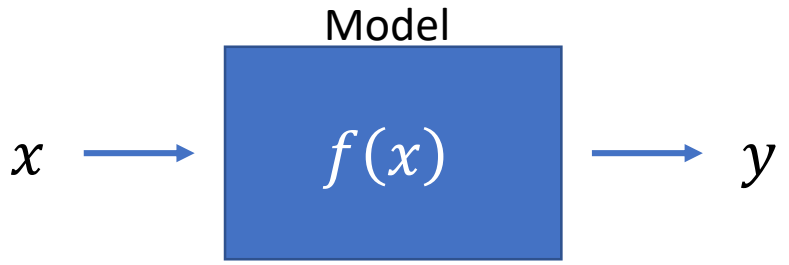


$x \to y$

$(x_1, x_2, \dots, x_n) \to y$

$(x_1, x_2, \dots, x_n) \to (y_1, y_2, \dots, y_k)$

Supervised



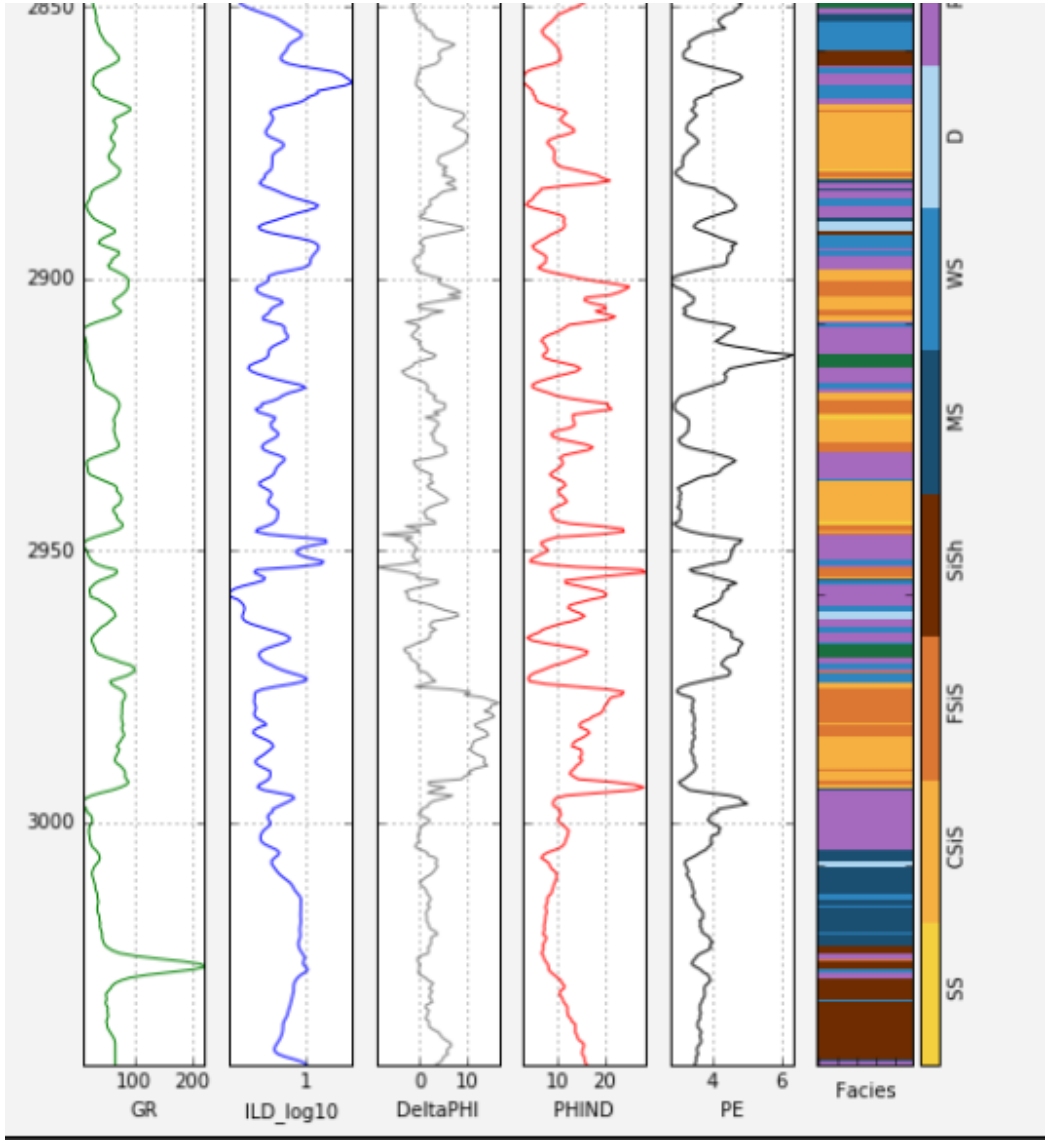
$f: X \to Y$

Unsupervised

ID	x_1	…	x_n	Category
1	3.532		A	Catx
2	7.234		H	Caty
⋮	⋮		⋮	⋮

ID	Cat y
1	aaa
2	hhh
⋮	⋮

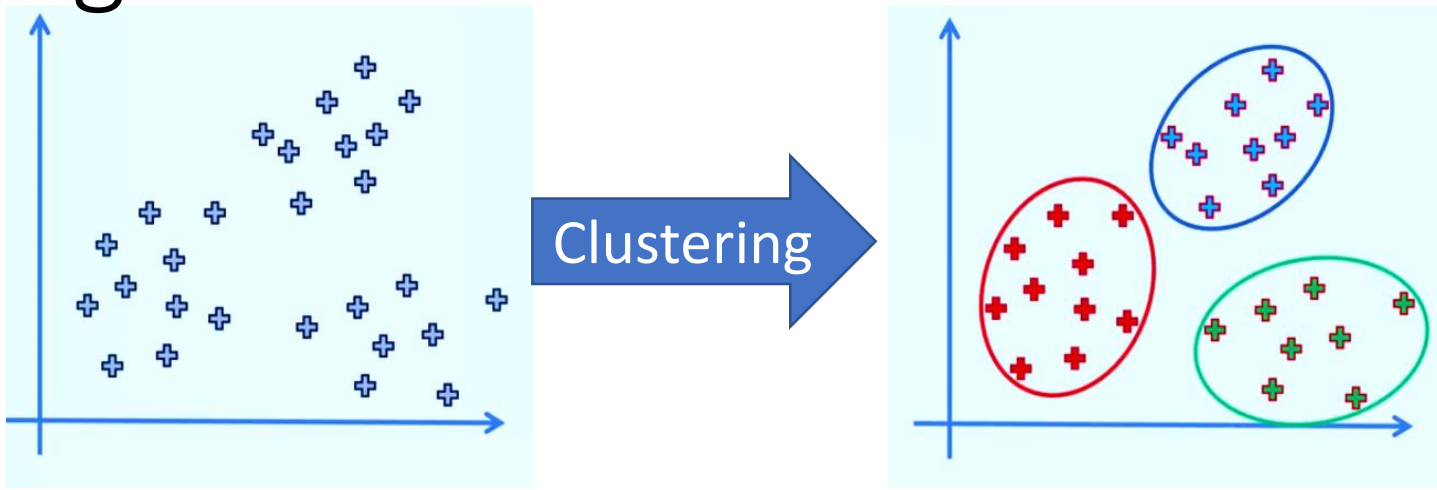
Supervised Vs Unsupervised learning



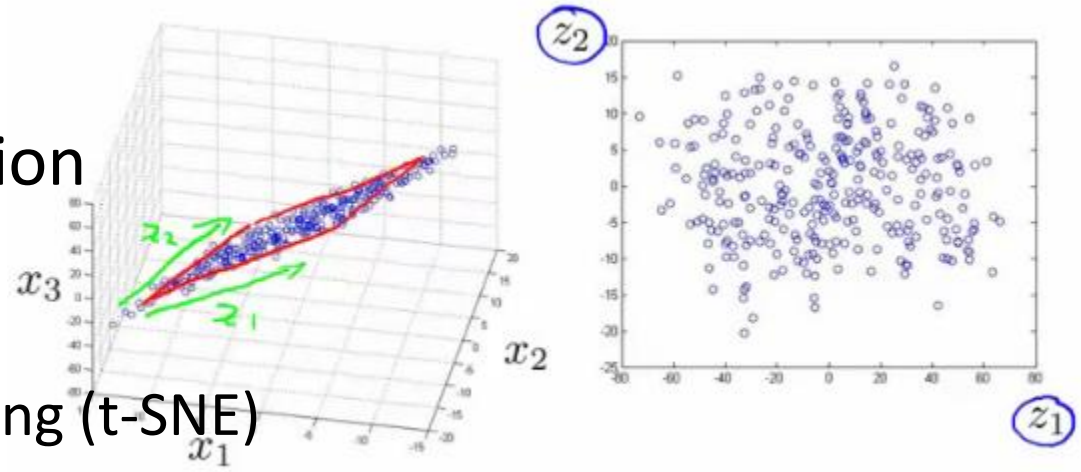
https://www.leouieda.com/talks/tgif2018.html

Unsupervised learning

- Clustering
  - K-means
  - DBSCAN
  - Hierarchical Cluster Analysis

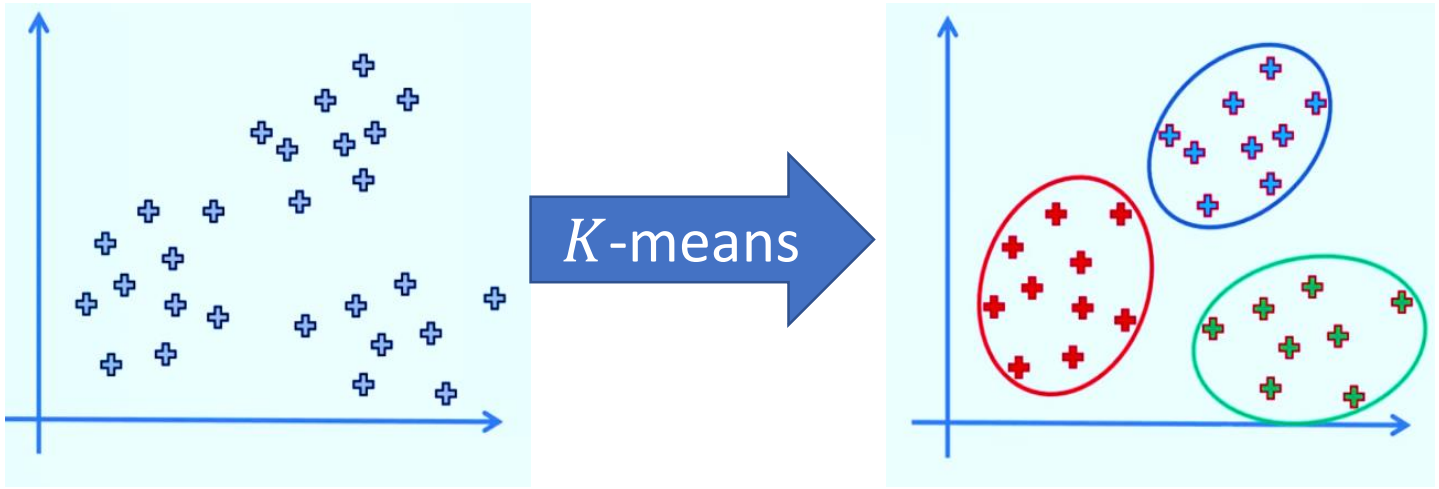


- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)



K-means

K-means



Assumptions

- K –clusters
- n instances

1. $C_1 \cup C_2 \cup \cdots \cup C_K = \{1, \ldots, n\}$
2. $C_i \cap C_j = \emptyset \; \forall \; i \neq j$

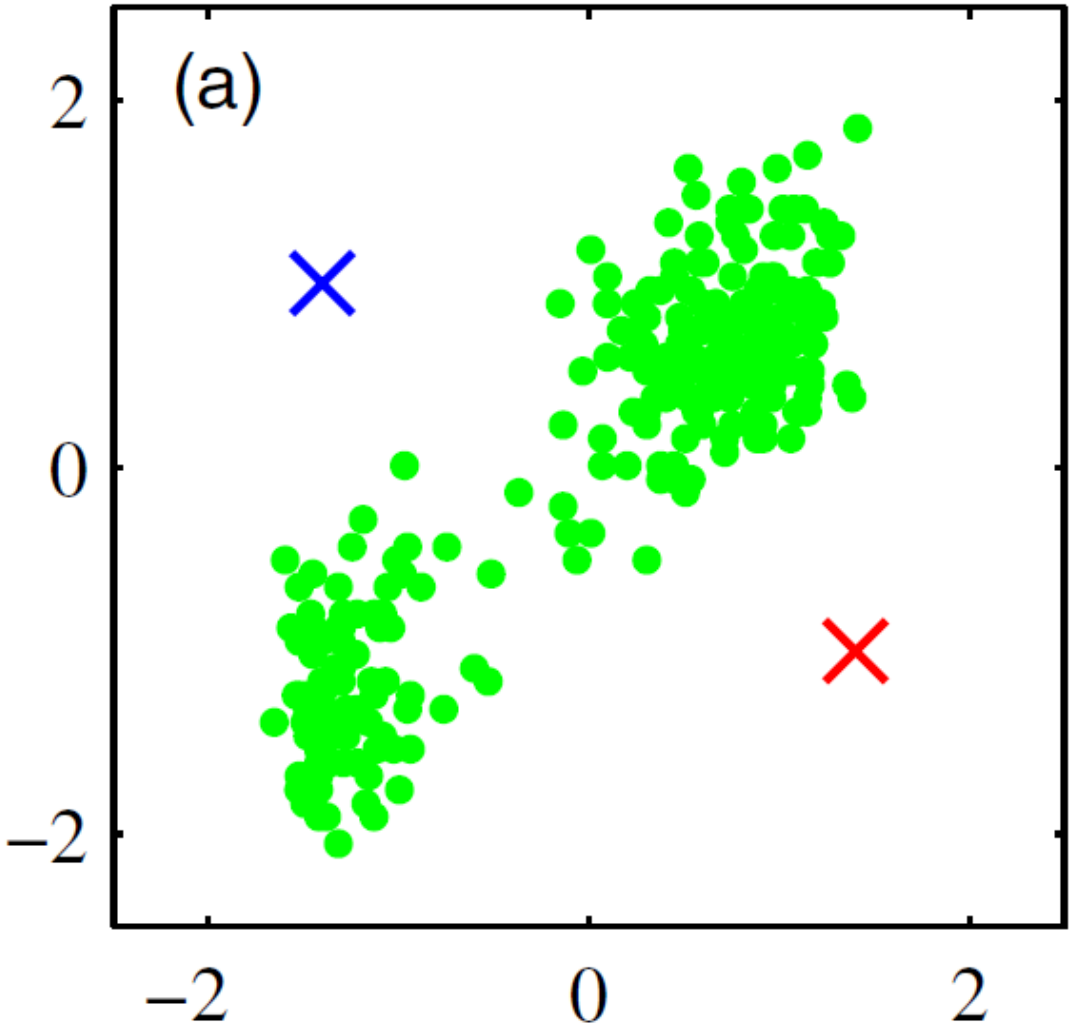
Requirements

Similarity or Dissimilarity (Distance) measure

Similarity vs Dissimilarity

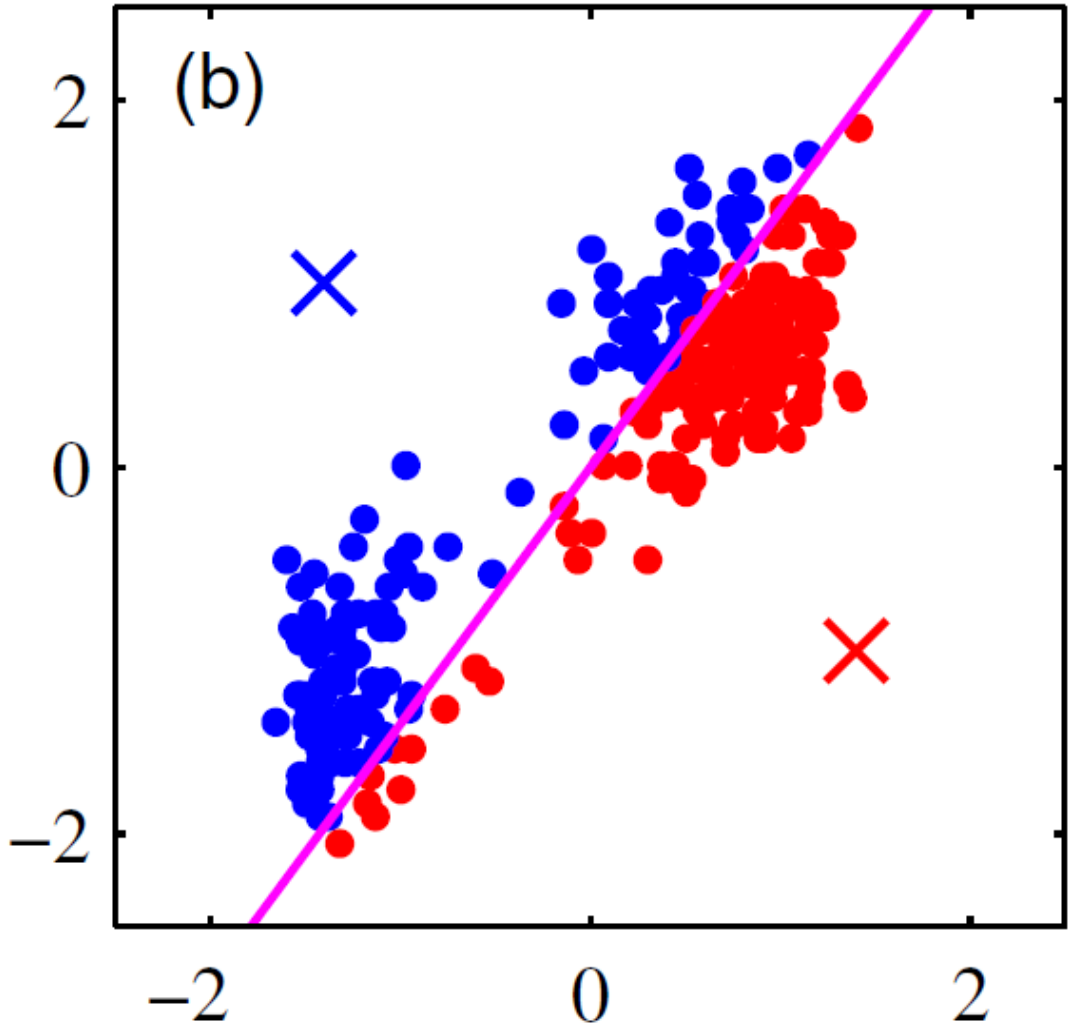
- The **similarity** between two objects is a numeral measure of the degree to which the two objects are alike. Consequently, similarities are higher for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1(complete similarity).
- The **dissimilarity** between two objects is the numerical measure of the degree to which the two objects are different. Dissimilarity is lower for more similar pairs of objects.
- Frequently, the term **distance** is used as a synonym for dissimilarity. Dissimilarities sometimes fall in the interval [0,1], but it is also common for them to range from 0 to ∞.

K-means algorithm



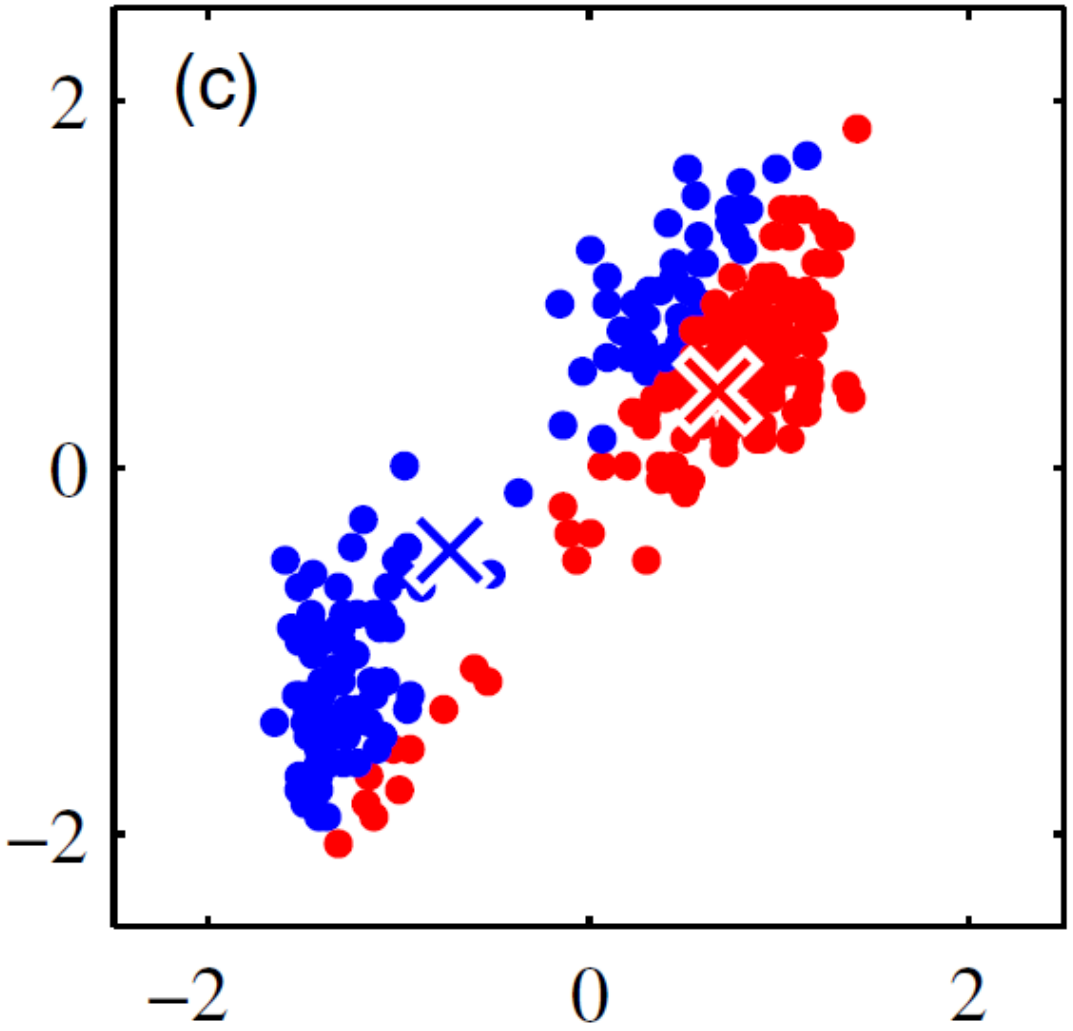
Bishop, 2007. Fig. 9.1

K-means algorithm



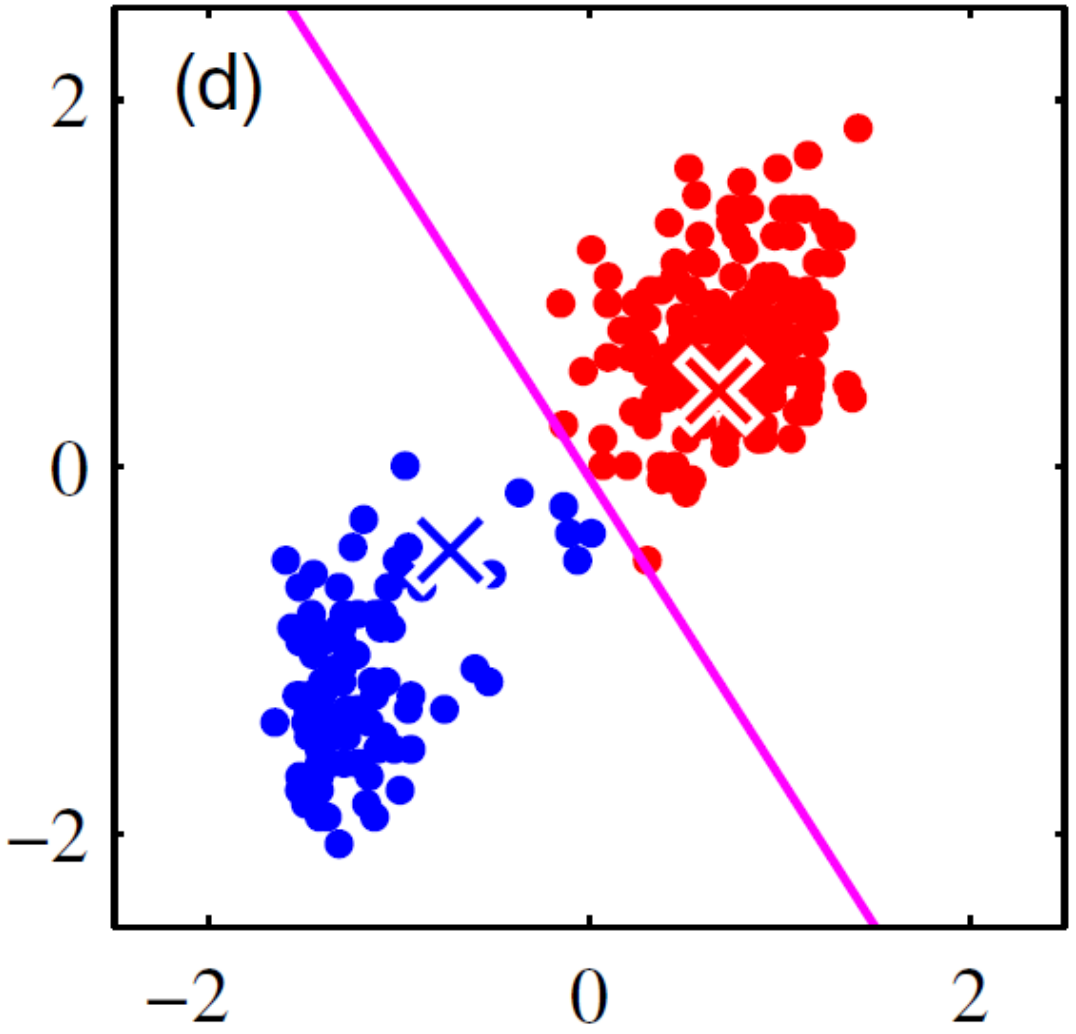
Bishop, 2007. Fig. 9.1

K-means algorithm



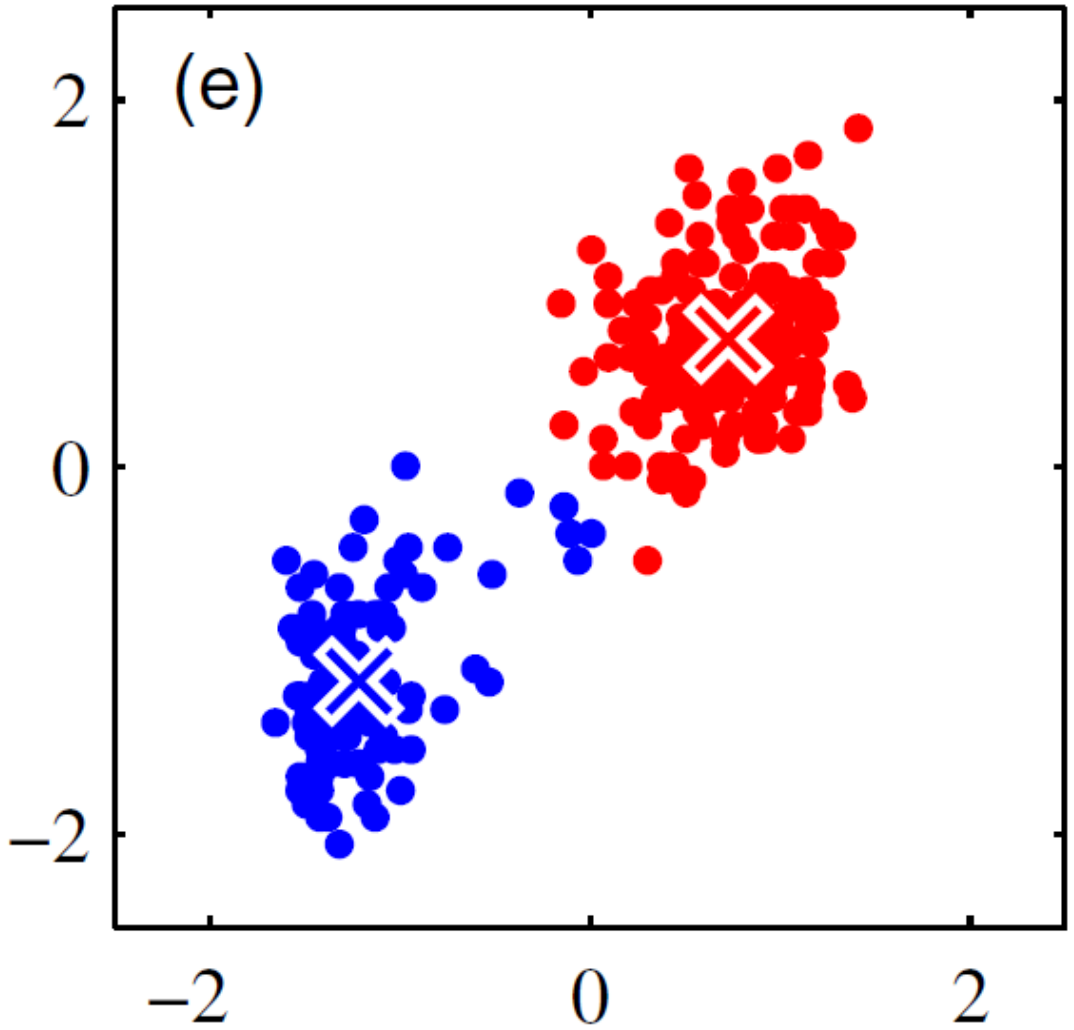
Bishop, 2007. Fig. 9.1

K-means algorithm



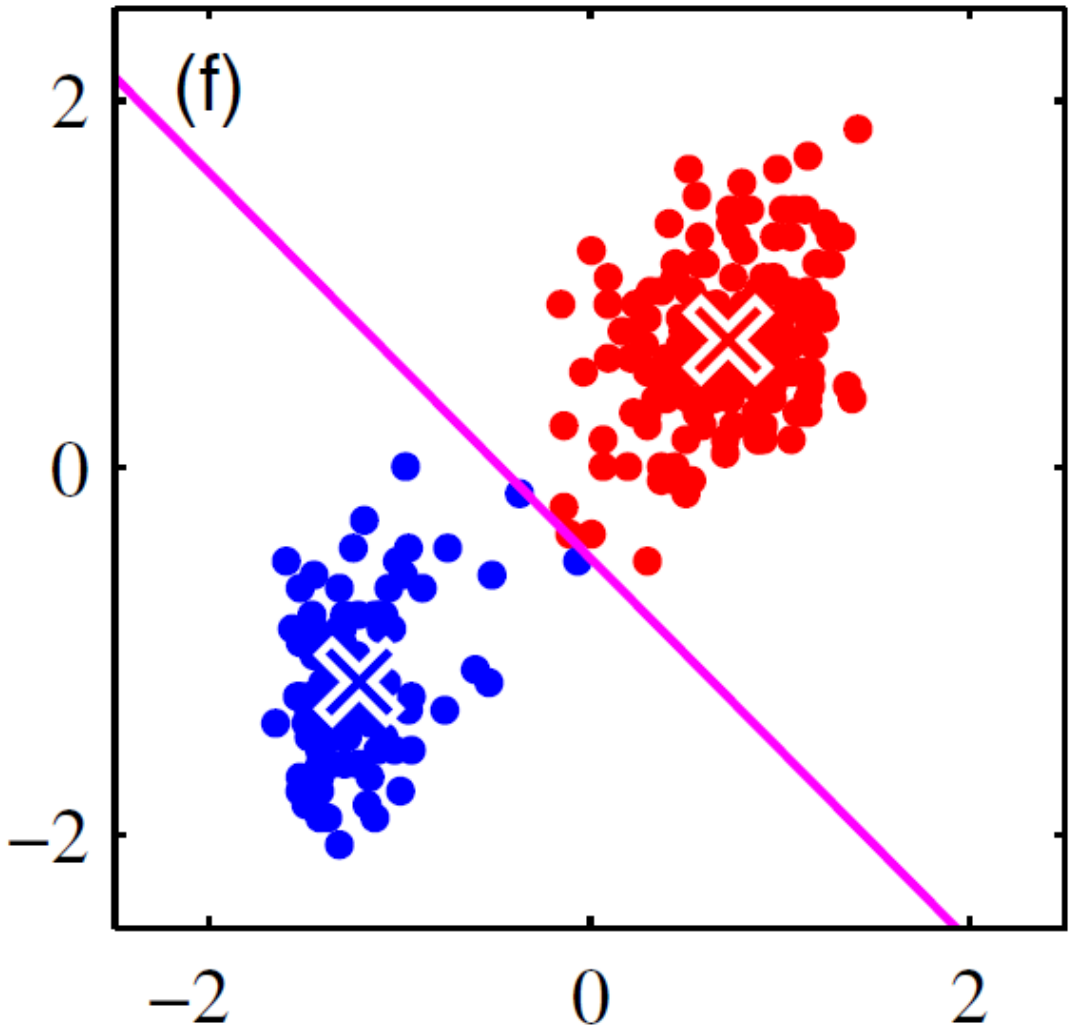
Bishop, 2007. Fig. 9.1

K-means algorithm



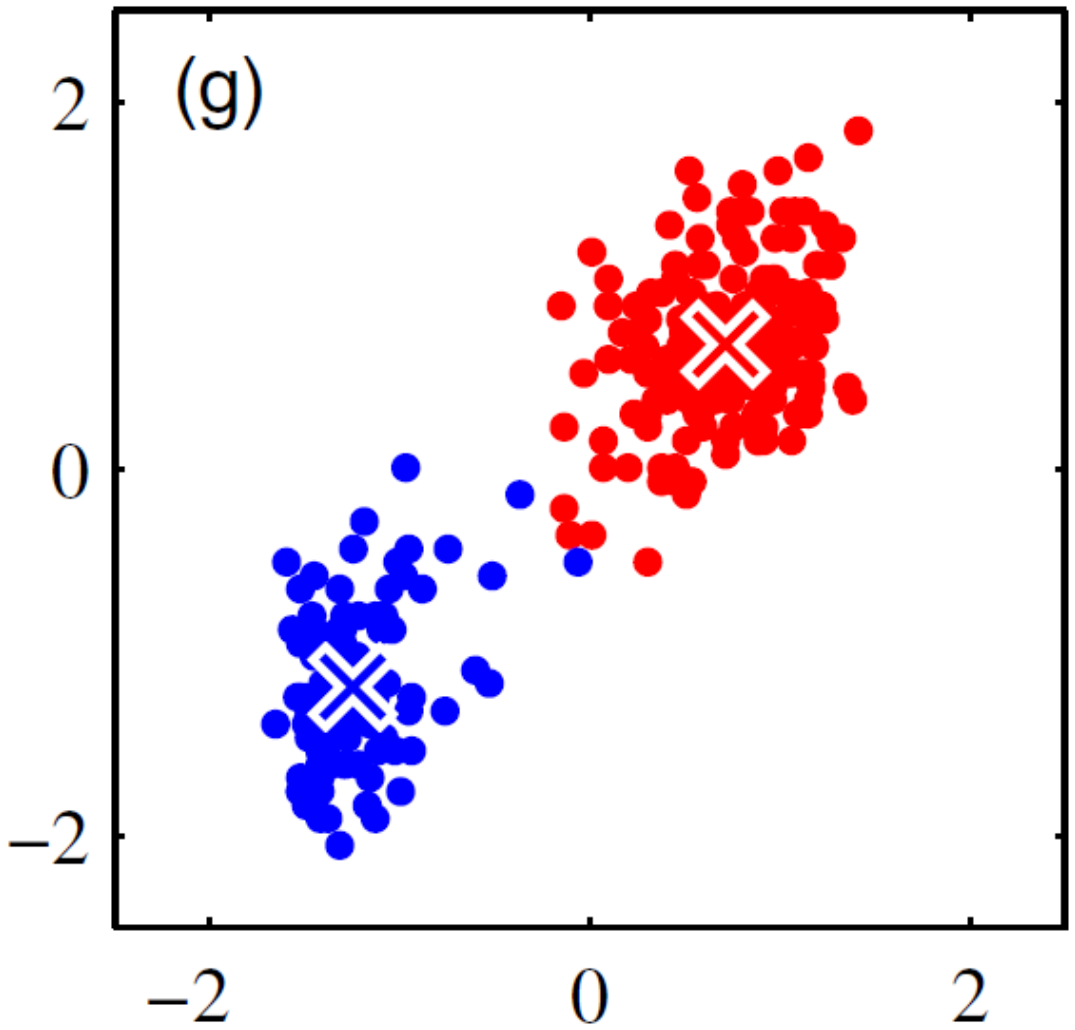
Bishop, 2007. Fig. 9.1

K-means algorithm



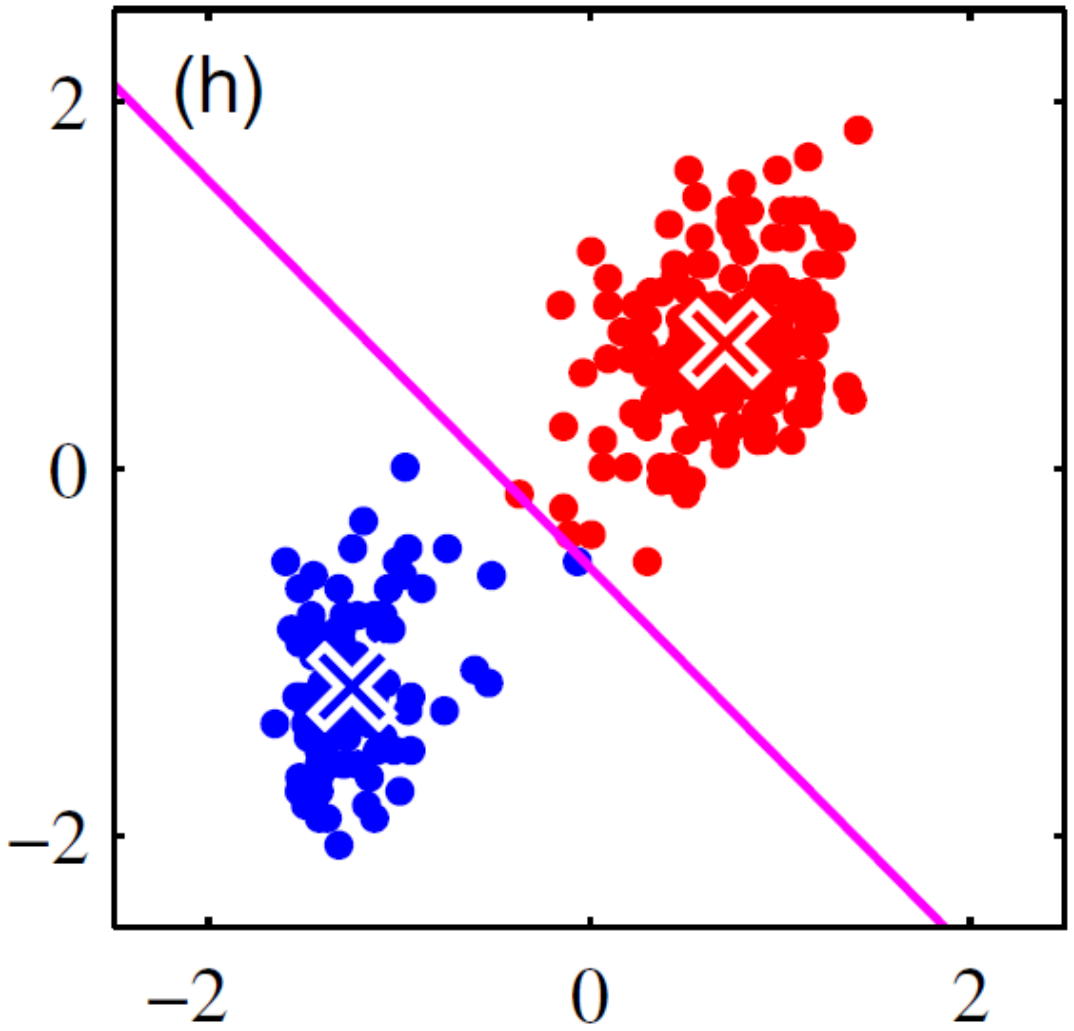
Bishop, 2007. Fig. 9.1

K-means algorithm



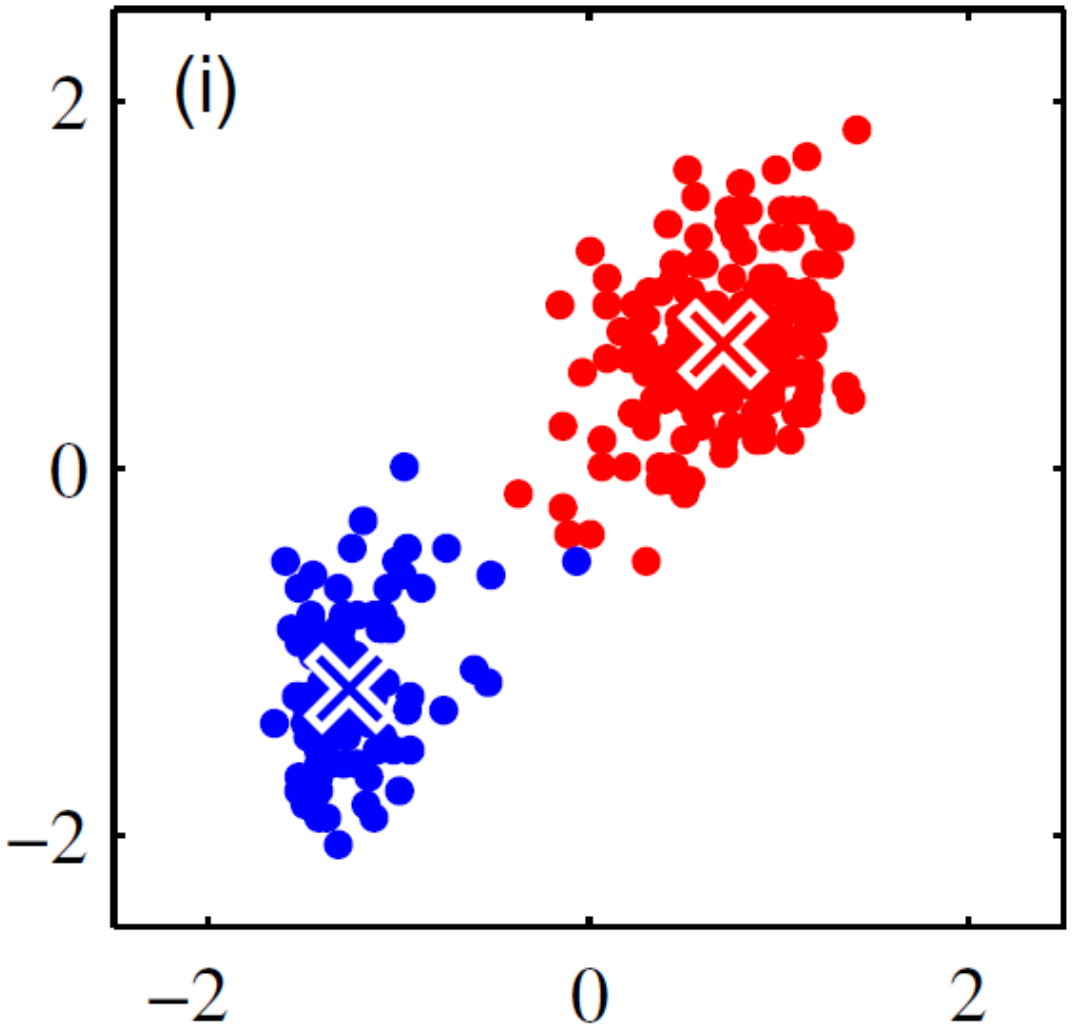
Bishop, 2007. Fig. 9.1

K-means algorithm



Bishop, 2007. Fig. 9.1

K-means algorithm



Bishop, 2007. Fig. 9.1

Exercise

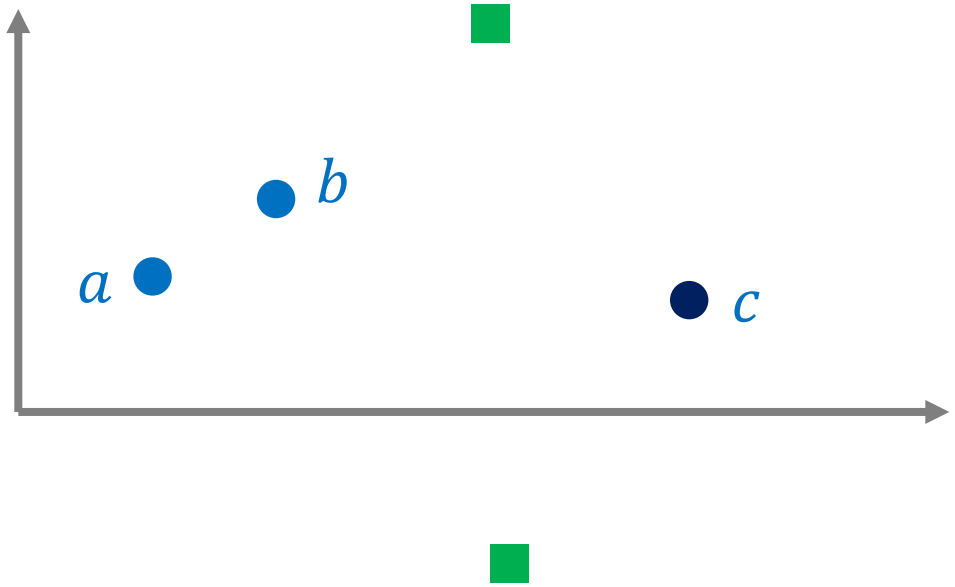
Data

	x	y
a	1	1
b	3	2
c	7	1

Initial Centroids

	x	y
c_1	5	3
c_2	5	1

$$\begin{pmatrix} \square & \square \\ \square & \square \end{pmatrix}$$



>_ Code

Homework assignment

- Generate a dataset with 3 groups and then use `sklearn.cluster.KMeans()` to get clusters of the dataset.

https://blog.floydhub.com/introduction-to-k-means-clustering-in-python-with-scikit-learn/