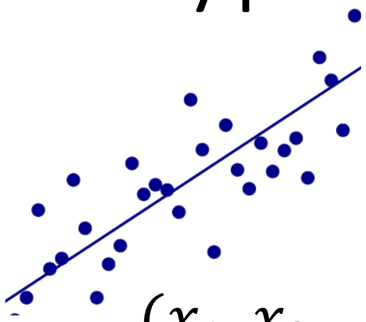
Supervised vs Unsupervised Learning

By Francisco Mendoza

mentofran@gmail.com

Type of problems, data types



$$x \rightarrow y$$

$$(x_1, x_2, \ldots, x_n) \rightarrow y$$

$$(x_1, x_2, \ldots, x_n) \rightarrow (y_1, y_2, \ldots, y_k)$$

Supervised

Model

$$x \longrightarrow \boxed{f(x)} \longrightarrow y$$

$$f: X \rightarrow Y$$

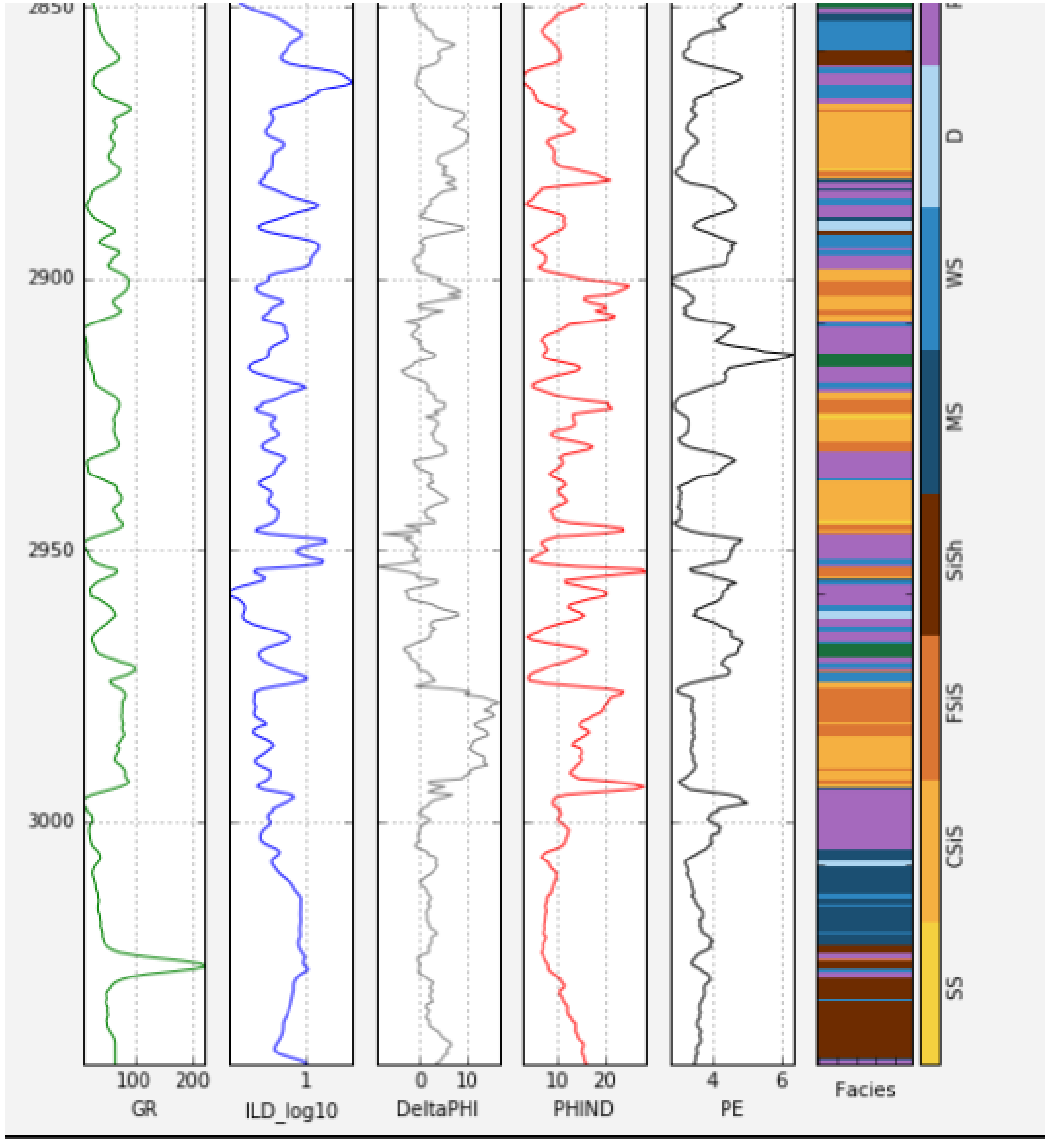
Unsupervised

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

$\longrightarrow$

ID	Cat y
1	aaa
2	hhh
⋮	⋮

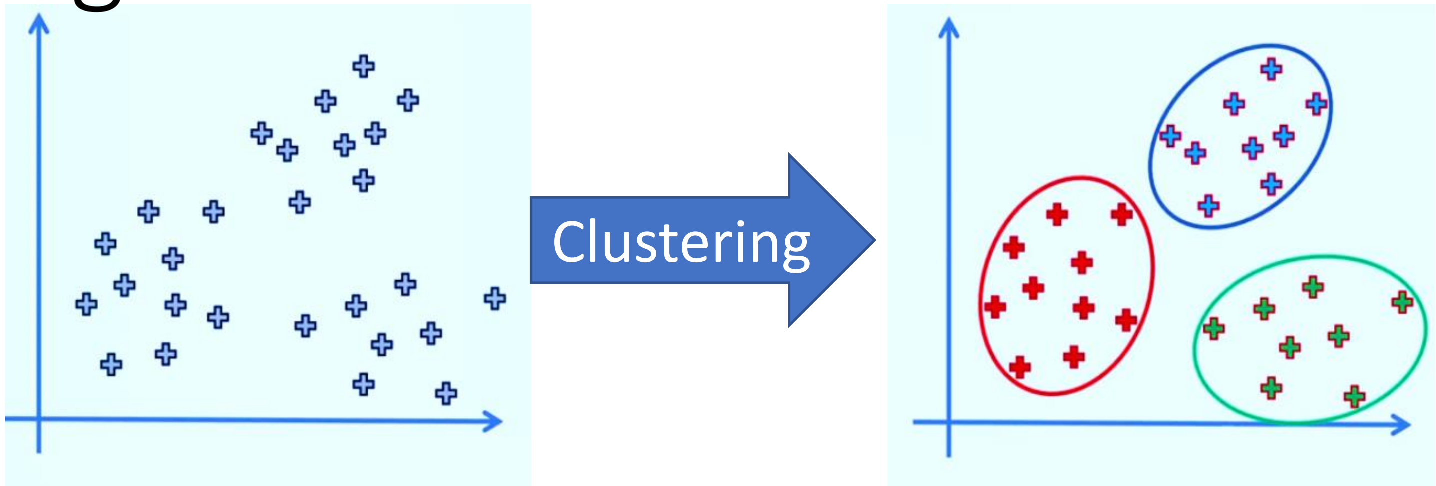
Supervised Vs Unsupervised learning



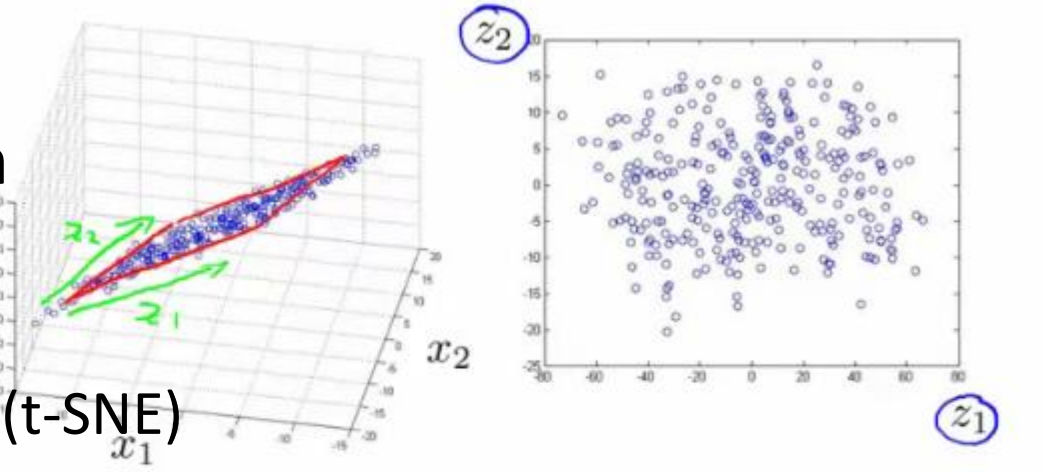
https://www.leouieda.com/talks/tgif2018.html

Unsupervised learning

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

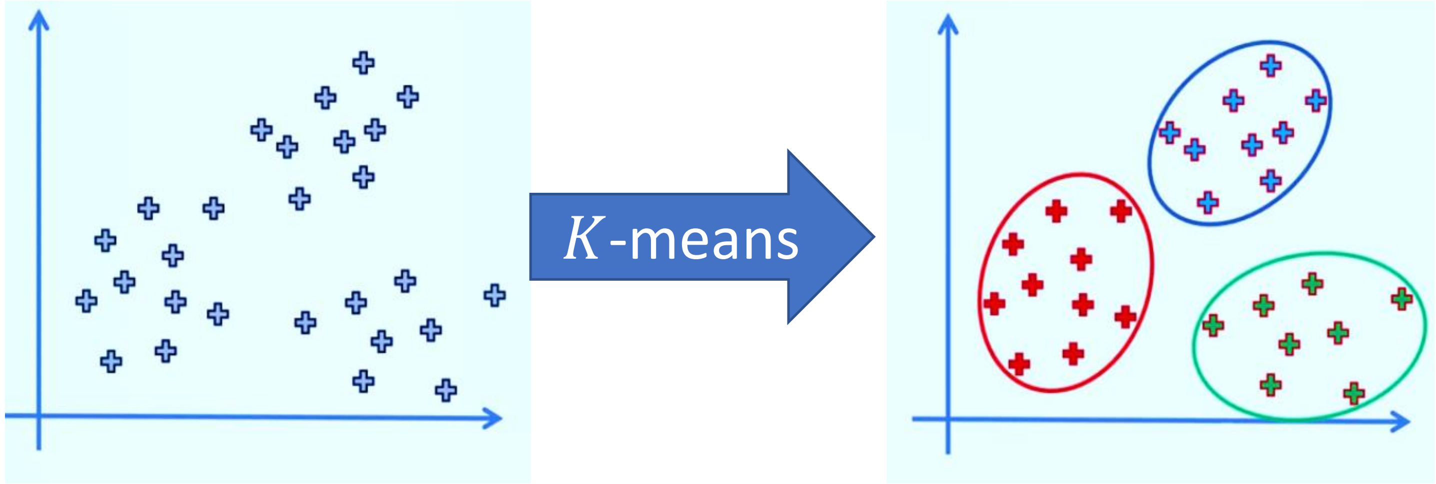


- Visualization and dimensionality reduction
  - Principal Component Analysis (PCA)
  - Locally-Linear Embedding (LLE)
  - t-distributed Stochastic Neighbor Embedding (t-SNE)



K-means

K-means



Assumptions

- K –clusters
- n instances

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

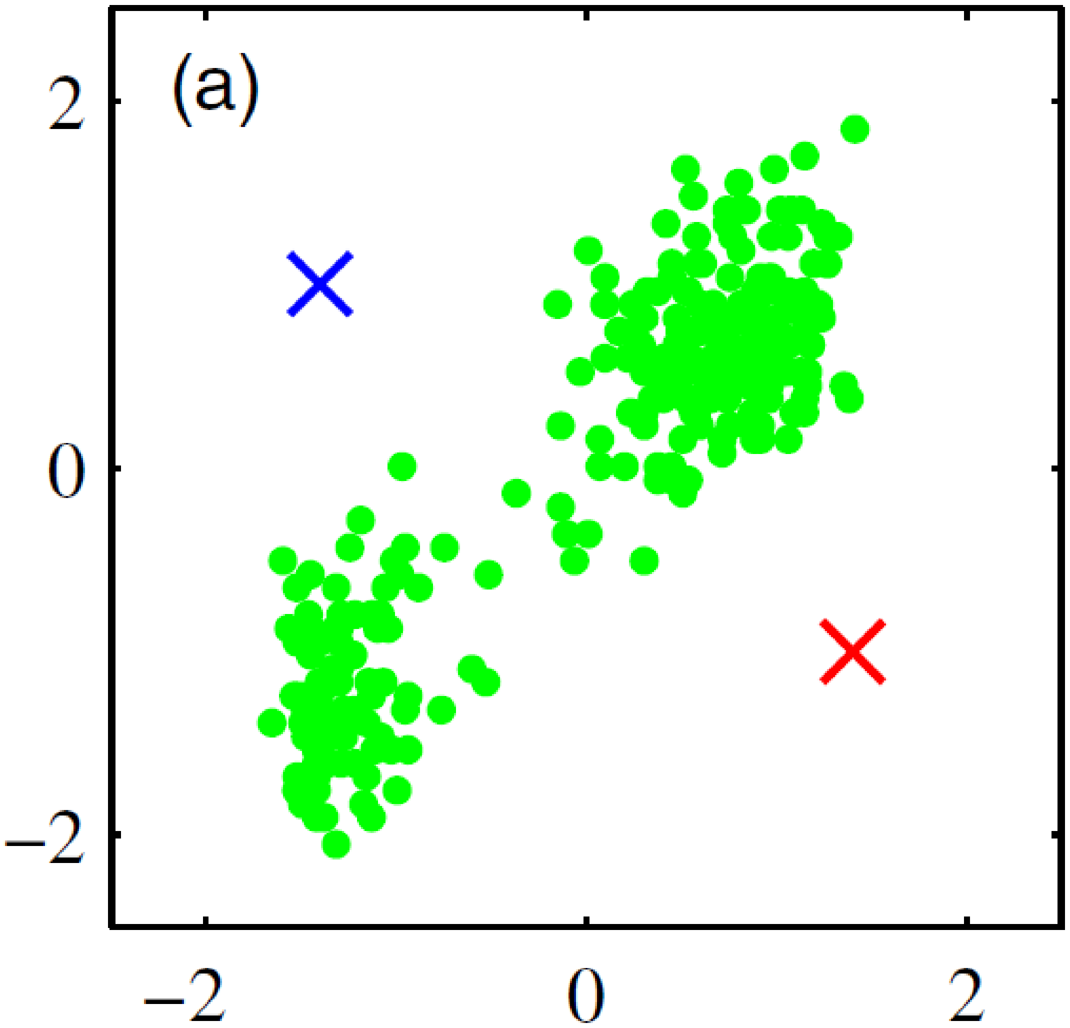
Requirements

Similarity or Dissimilarity (Distance) measure

Similarity vs Dissimilarity

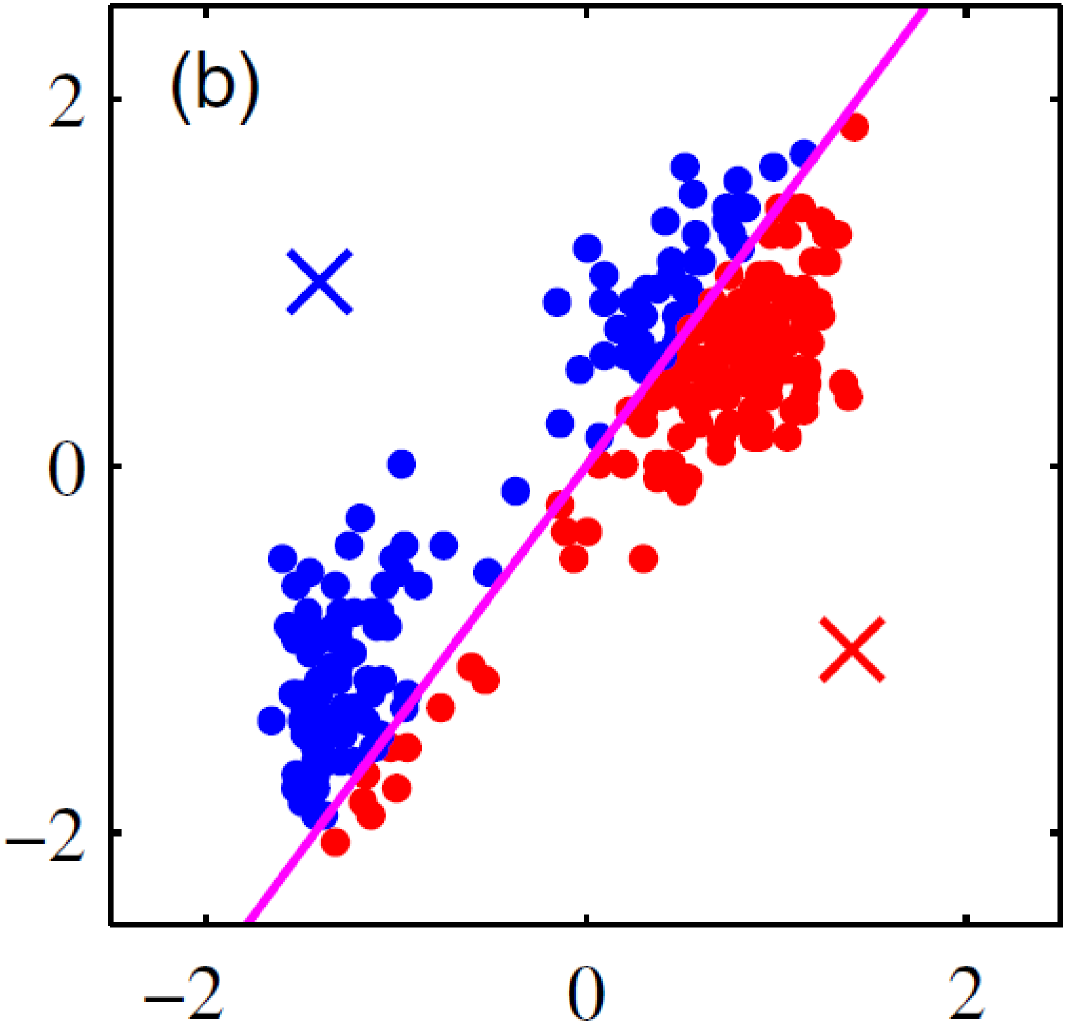
- The **similarity** between two objects is a numeral measure of the degree to which the two objects are alike. Consequently, similarities are higher for pairs of objects that are more alike. Similarities are usually non-negative and are often between 0 (no similarity) and 1(complete similarity).
- The **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



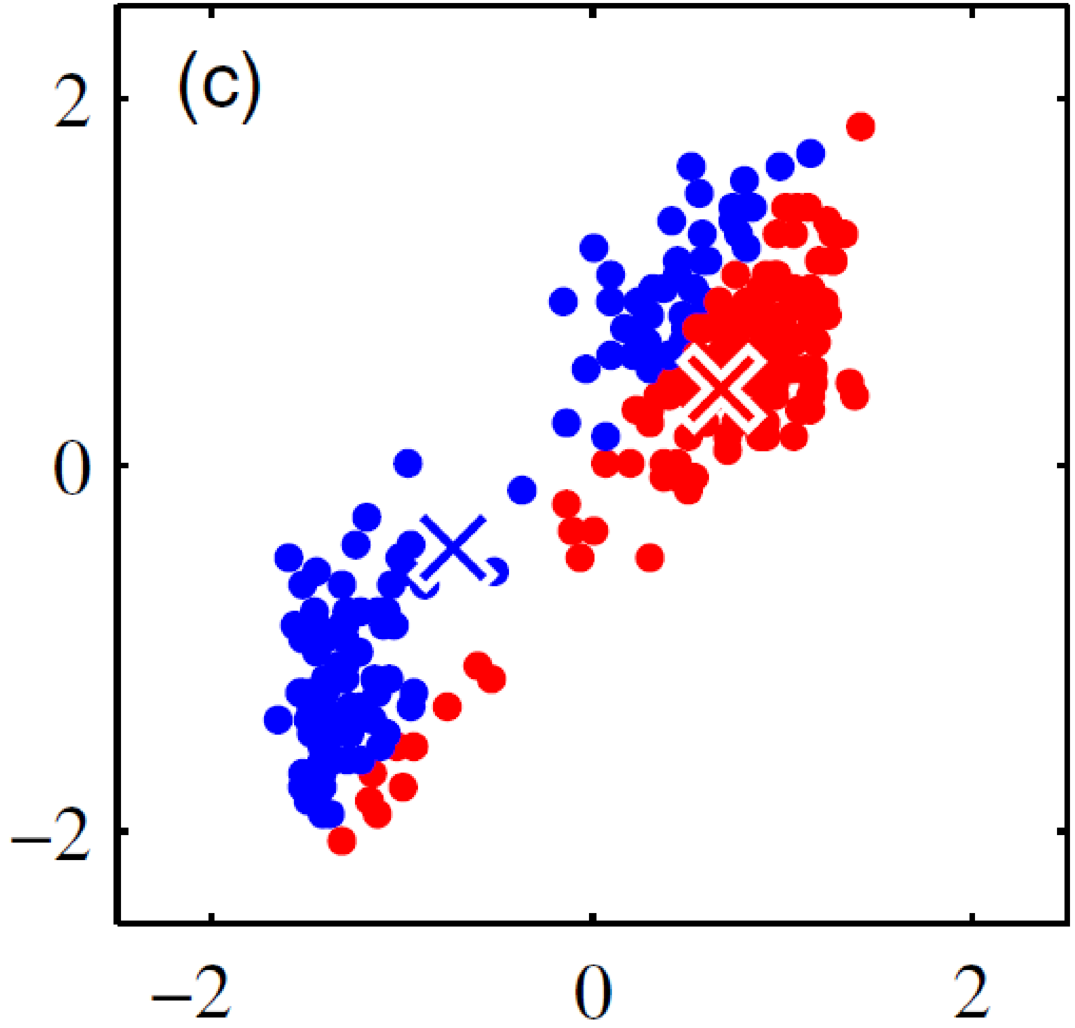
Bishop, 2007. Fig. 9.1

K-means algorithm



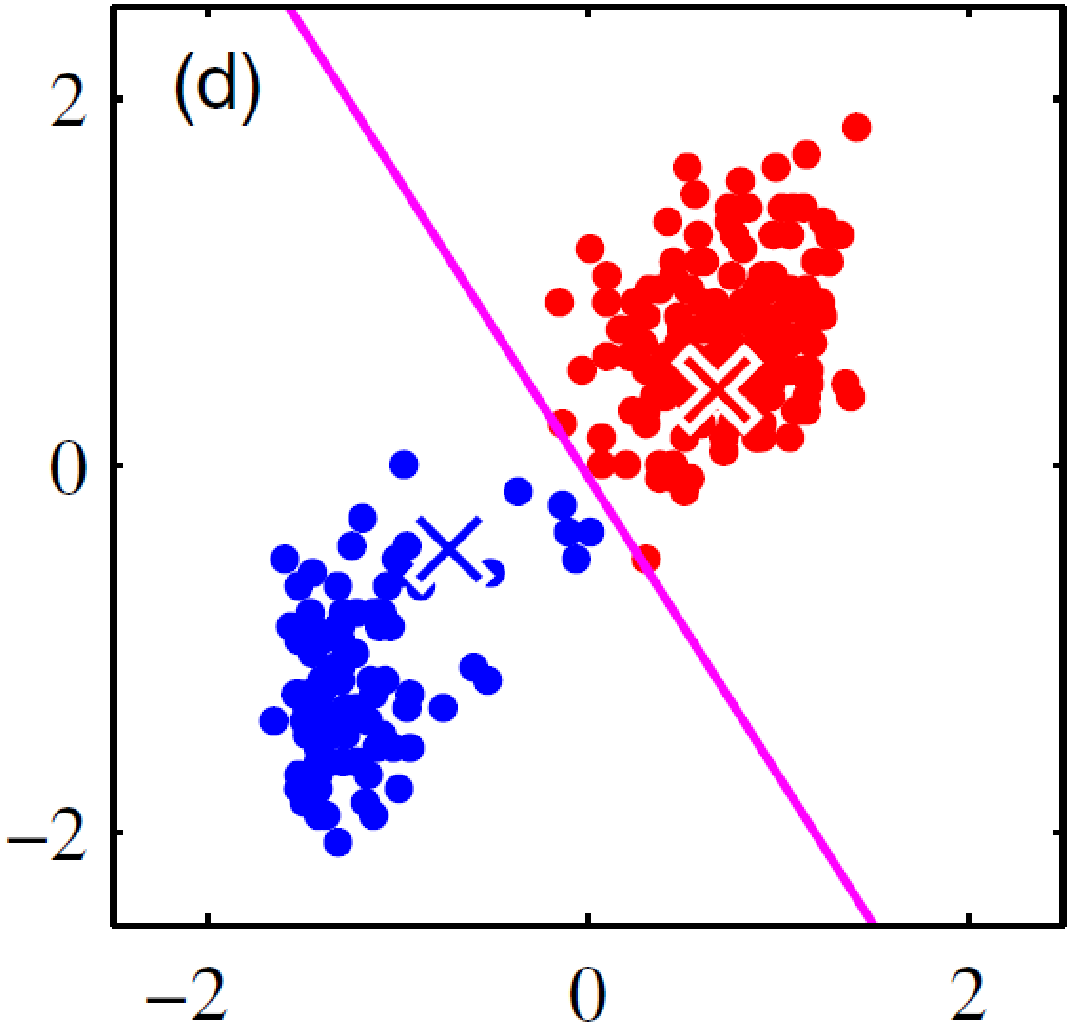
Bishop, 2007. Fig. 9.1

K-means algorithm



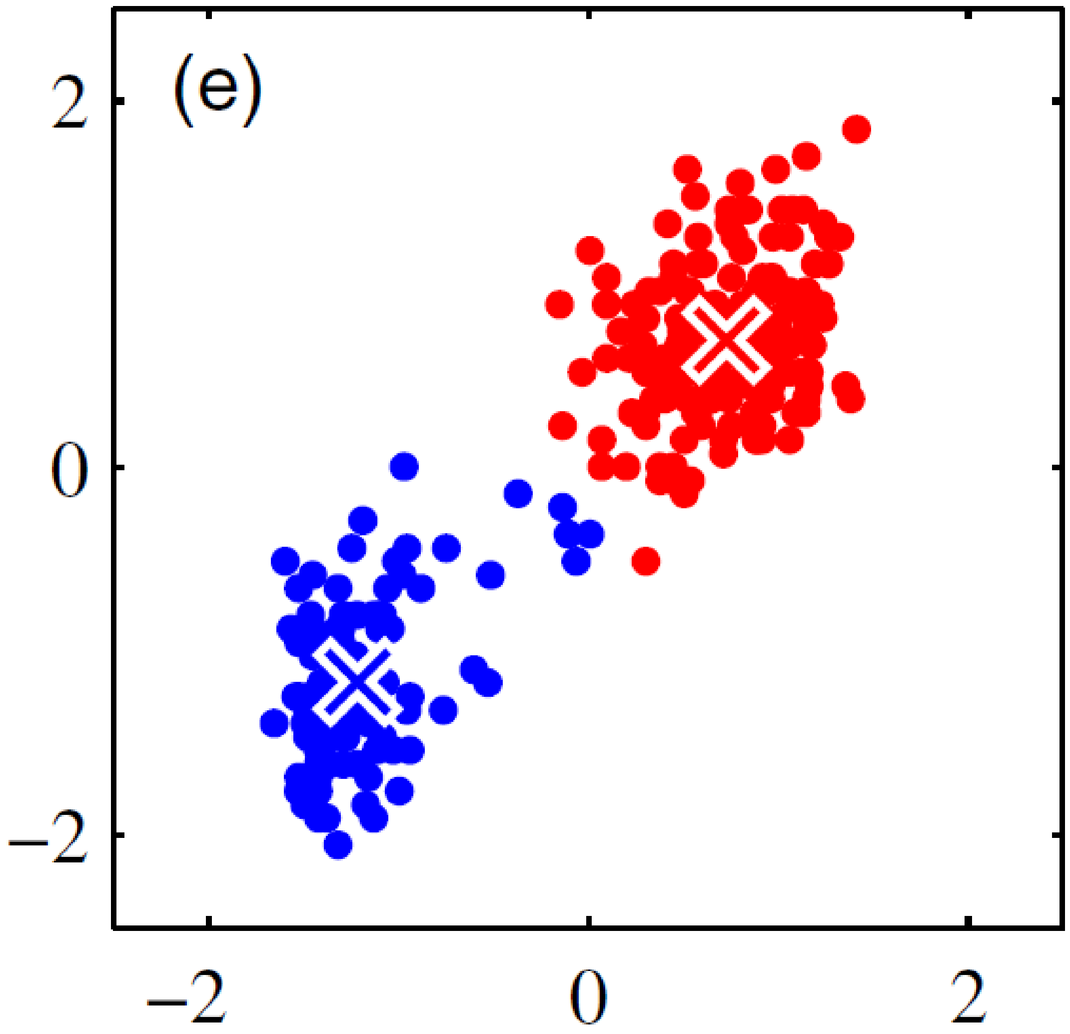
Bishop, 2007. Fig. 9.1

K-means algorithm



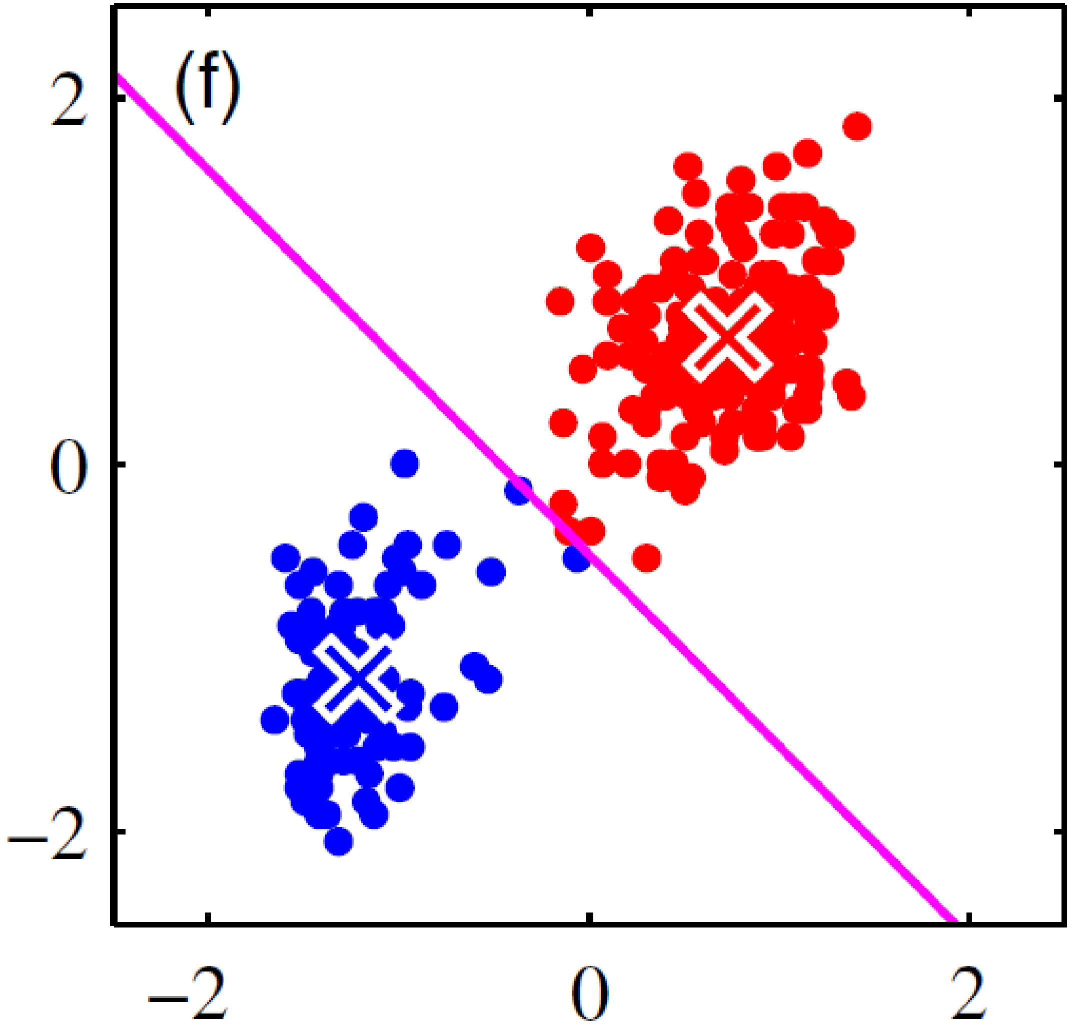
Bishop, 2007. Fig. 9.1

K-means algorithm



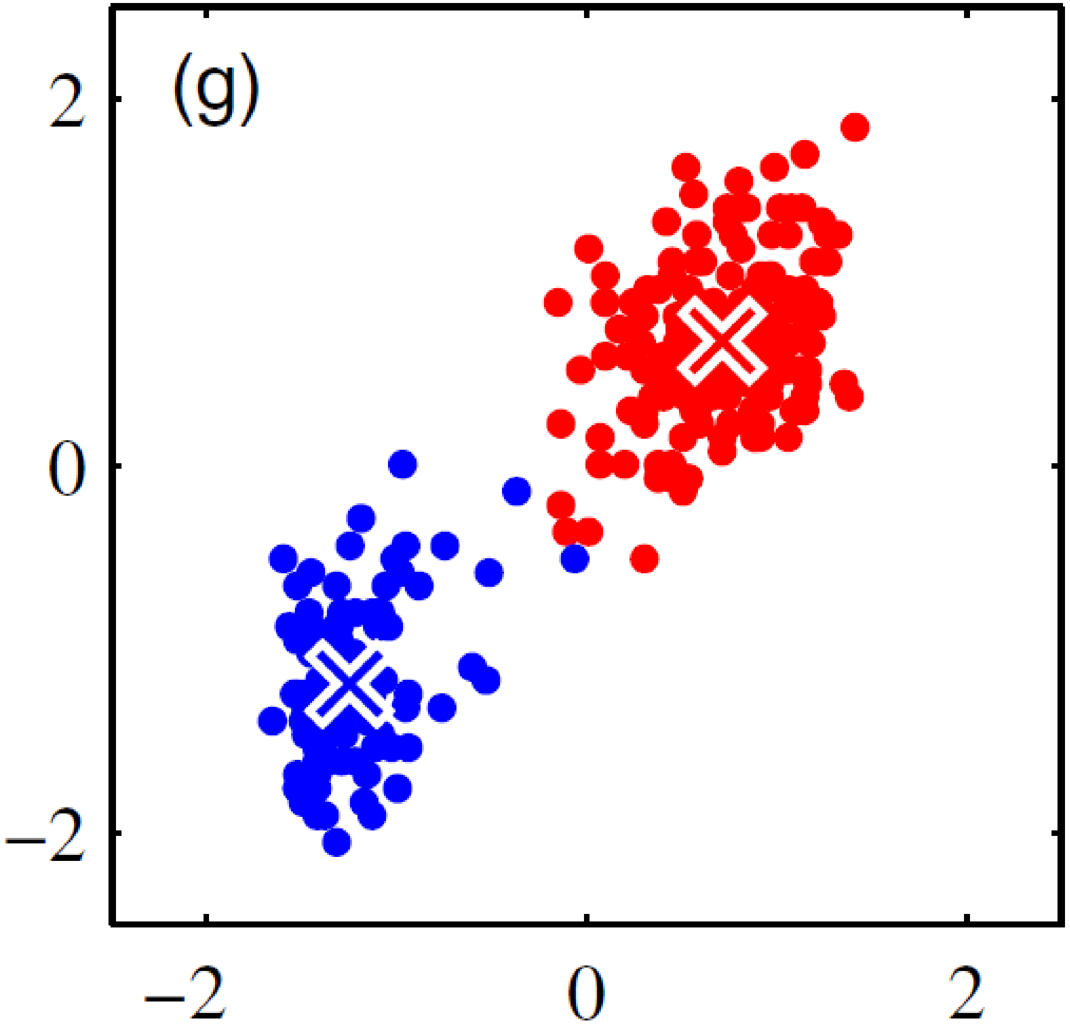
Bishop, 2007. Fig. 9.1

K-means algorithm



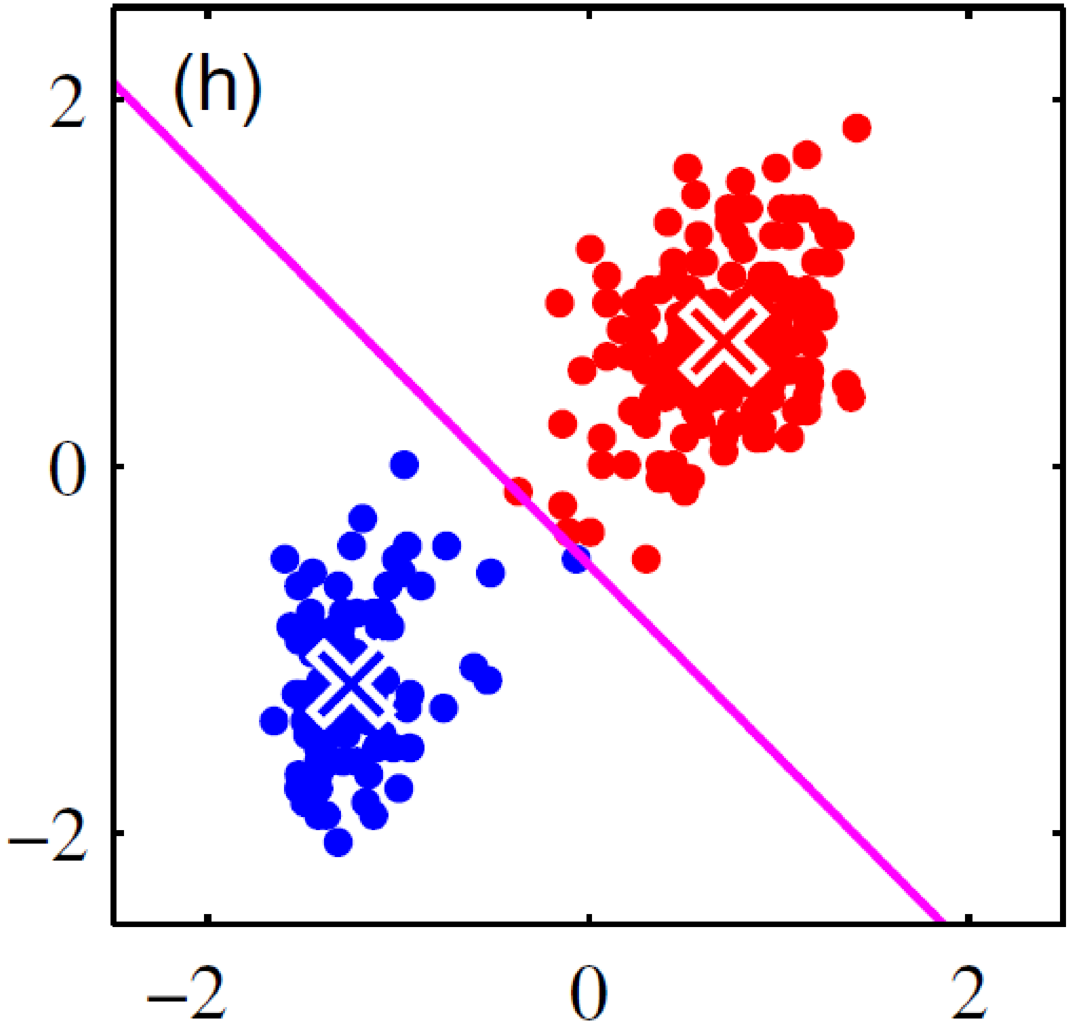
Bishop, 2007. Fig. 9.1

K-means algorithm



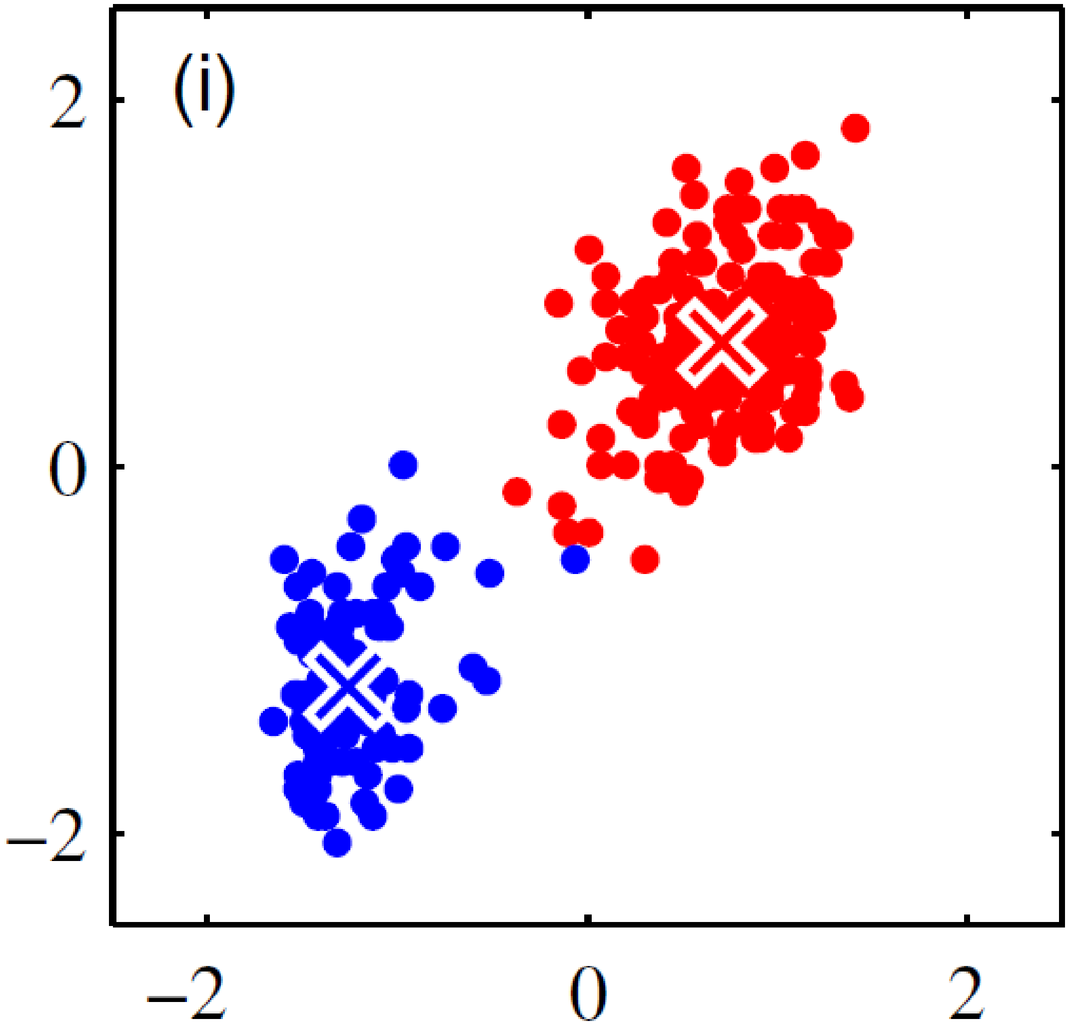
Bishop, 2007. Fig. 9.1

K-means algorithm



Bishop, 2007. Fig. 9.1

K-means algorithm



Bishop, 2007. Fig. 9.1

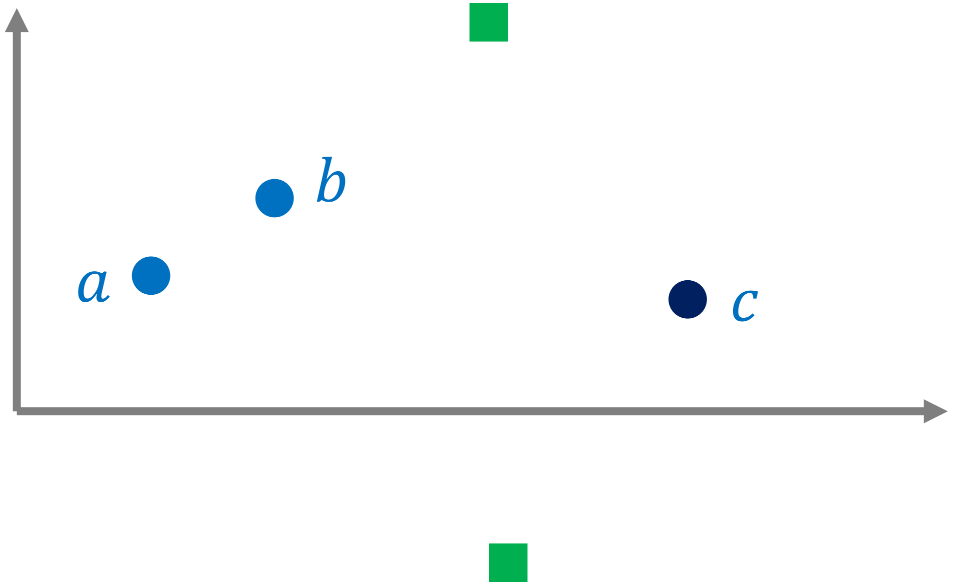
Exercise

Data

	x	y
a	1	1
b	3	2
c	7	1

Initial Centroids

	x	y
c_1	5	3
c_2	5	1



>_ Code

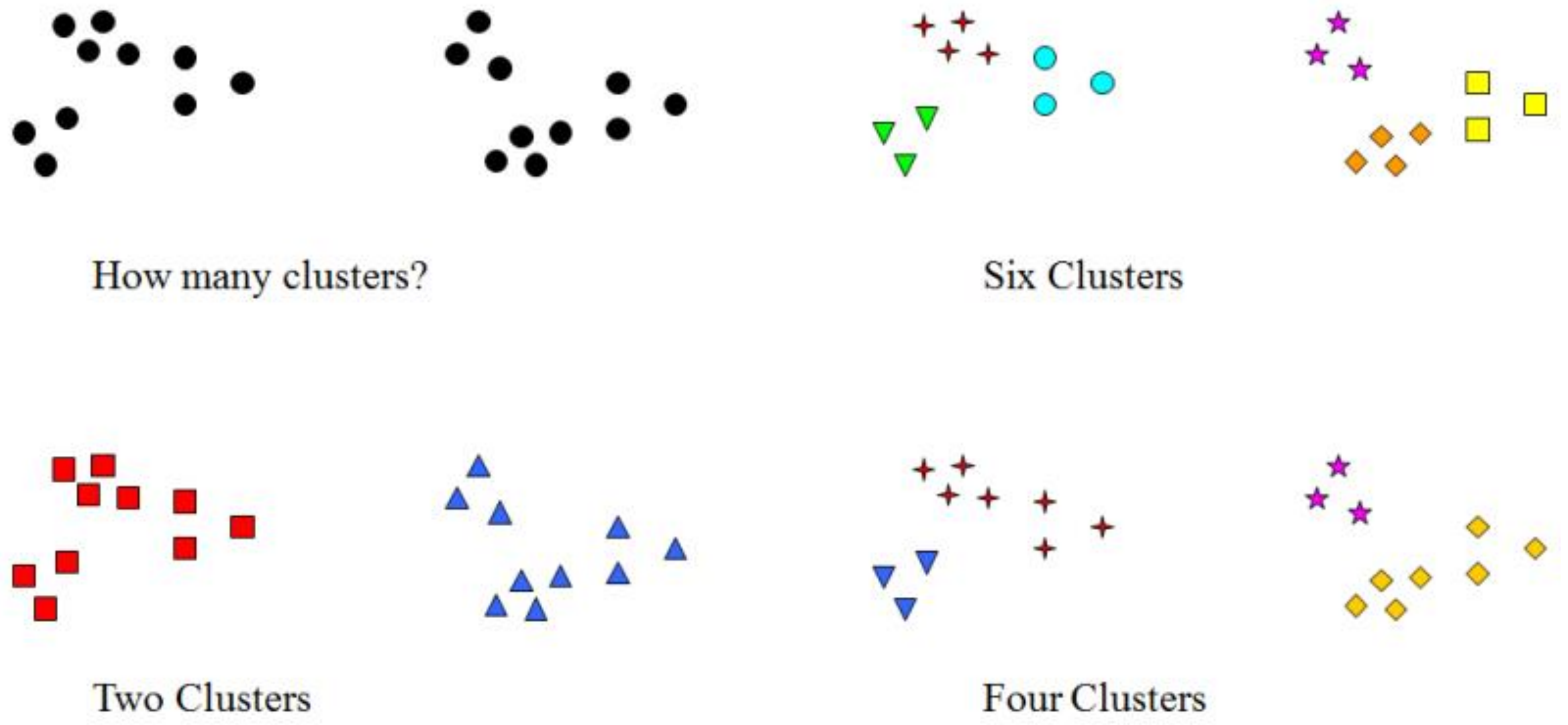
Homework assignment

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

- Plot the scatterplot identifying each cluster with a different color

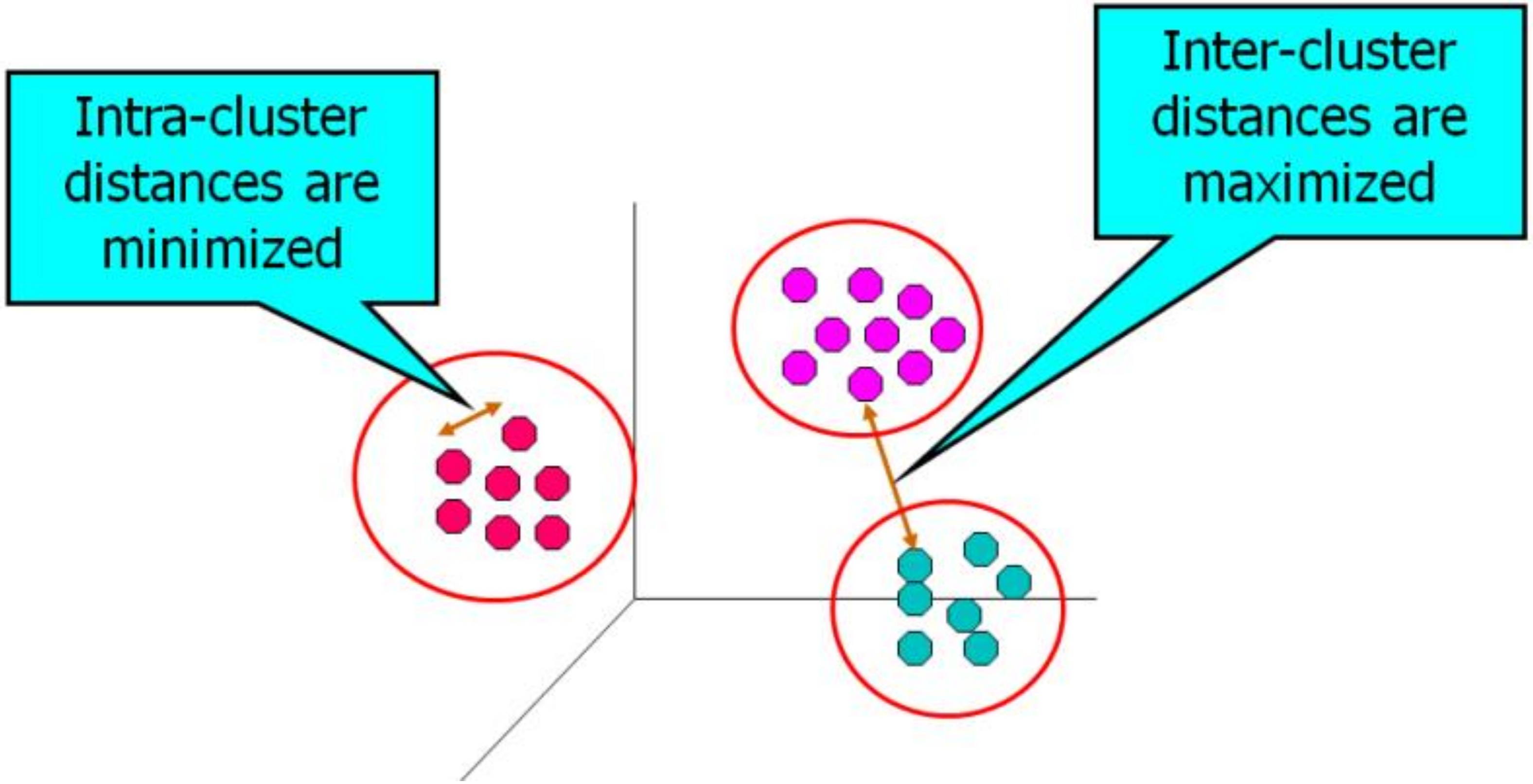
https://blog.floydhub.com/introduction-to-k-means-clustering-in-python-with-scikit-learn/

Choosing K. Silhouette coefficient/score



https://visibledata.wordpress.com/visible-data/cluster-analysis/

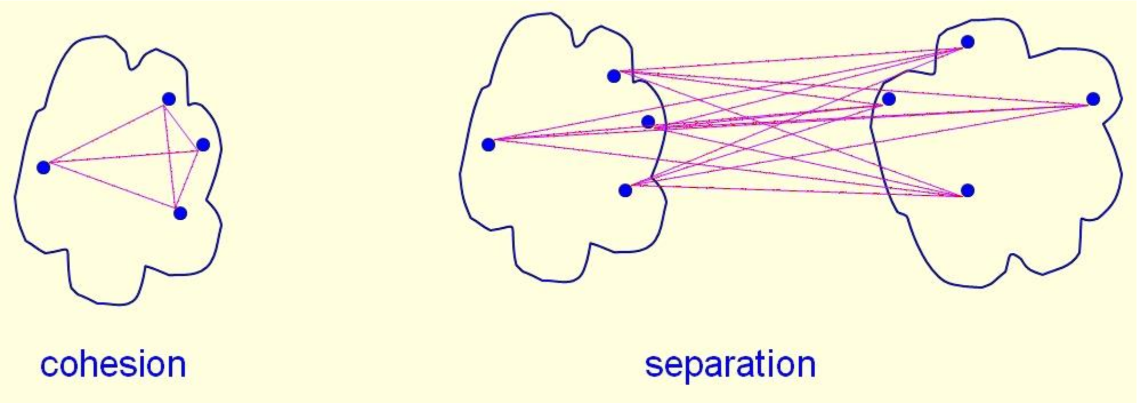
Choosing K. Silhouette coefficient/score



https://visibledata.wordpress.com/visible-data/cluster-analysis/

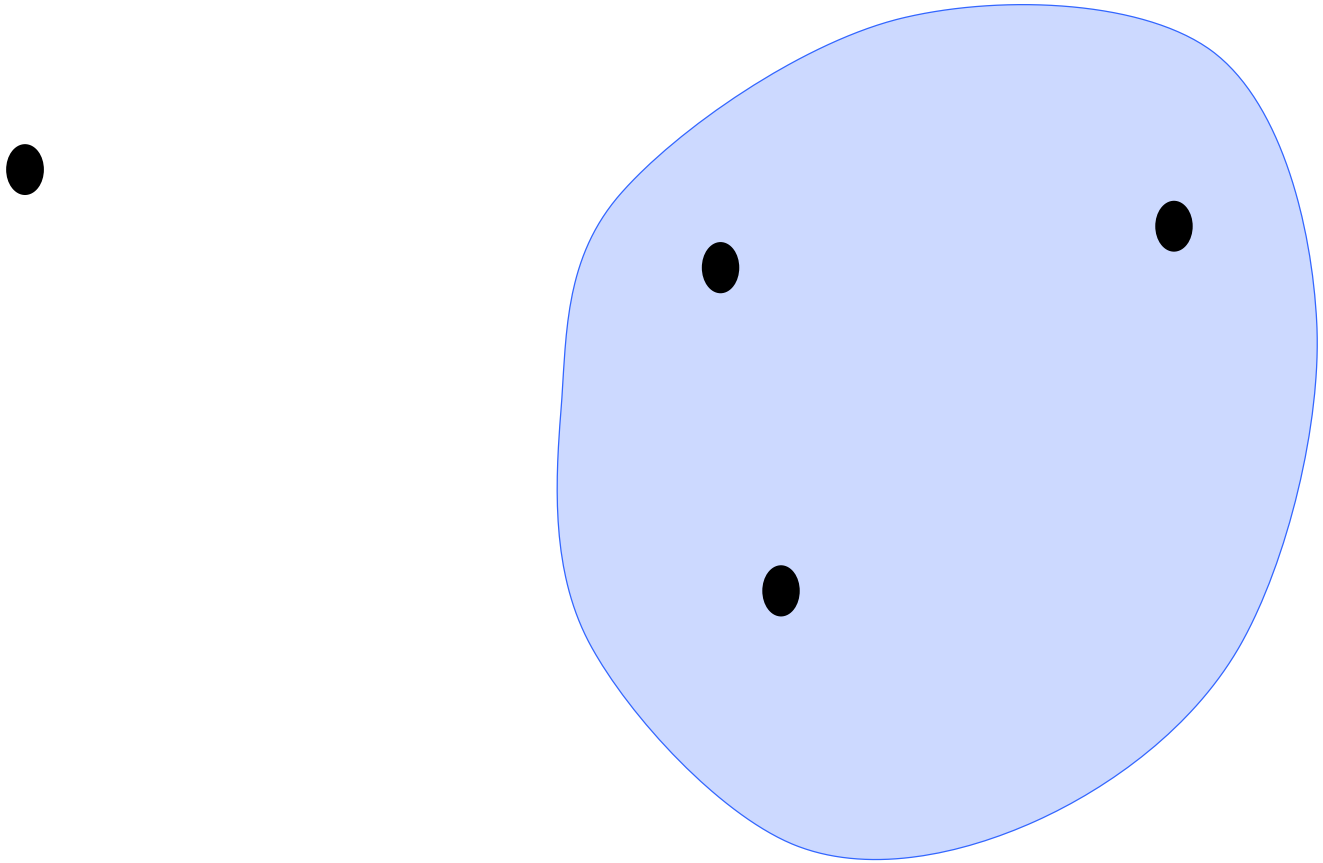
Cohesion and Separation

- Cluster cohesion
  - How tightly packed is a cluster
  - More cohesive clusters is more better
- Cluster separation
  - Distance between clusters
  - The more separation, the better
- Can we measure these things?
  - Yes



Cohesion (intra-cluster)

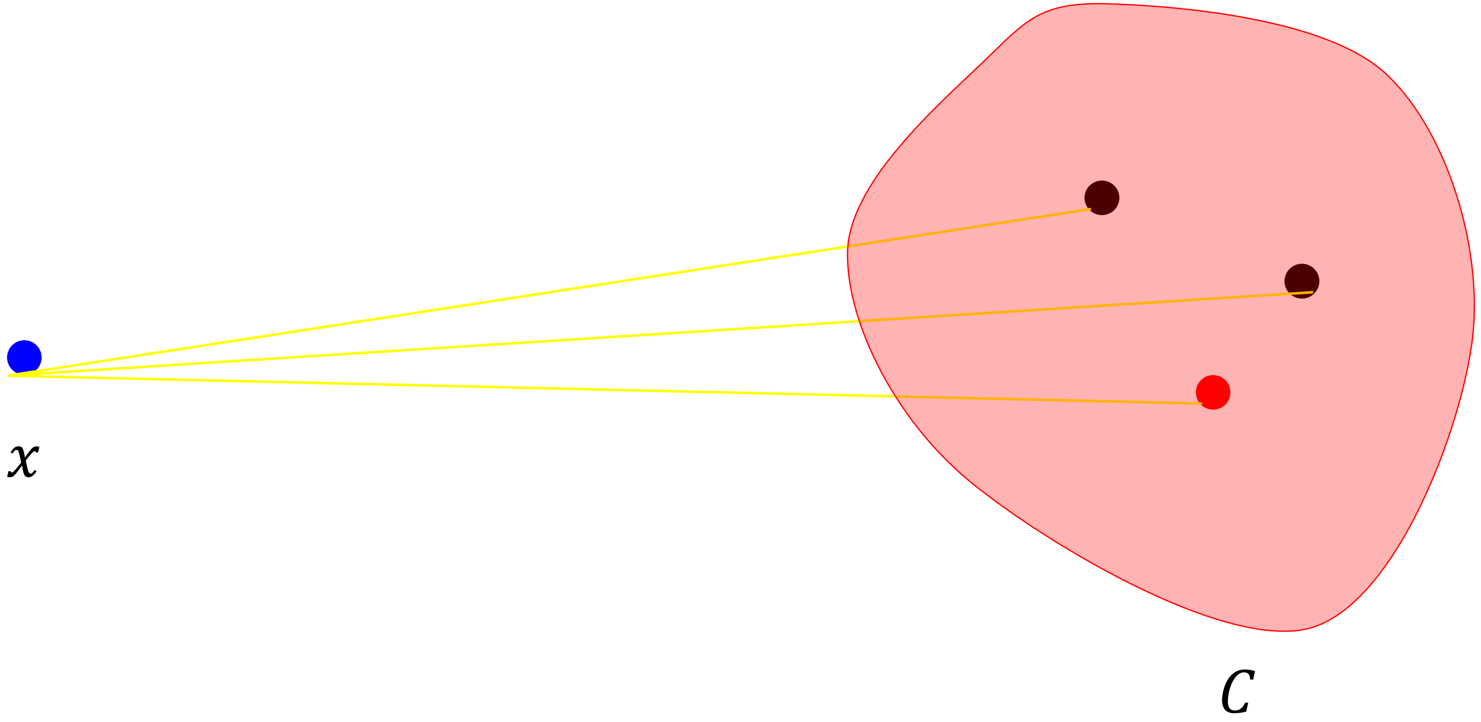
- For a data point x_i in the cluster C_i



$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

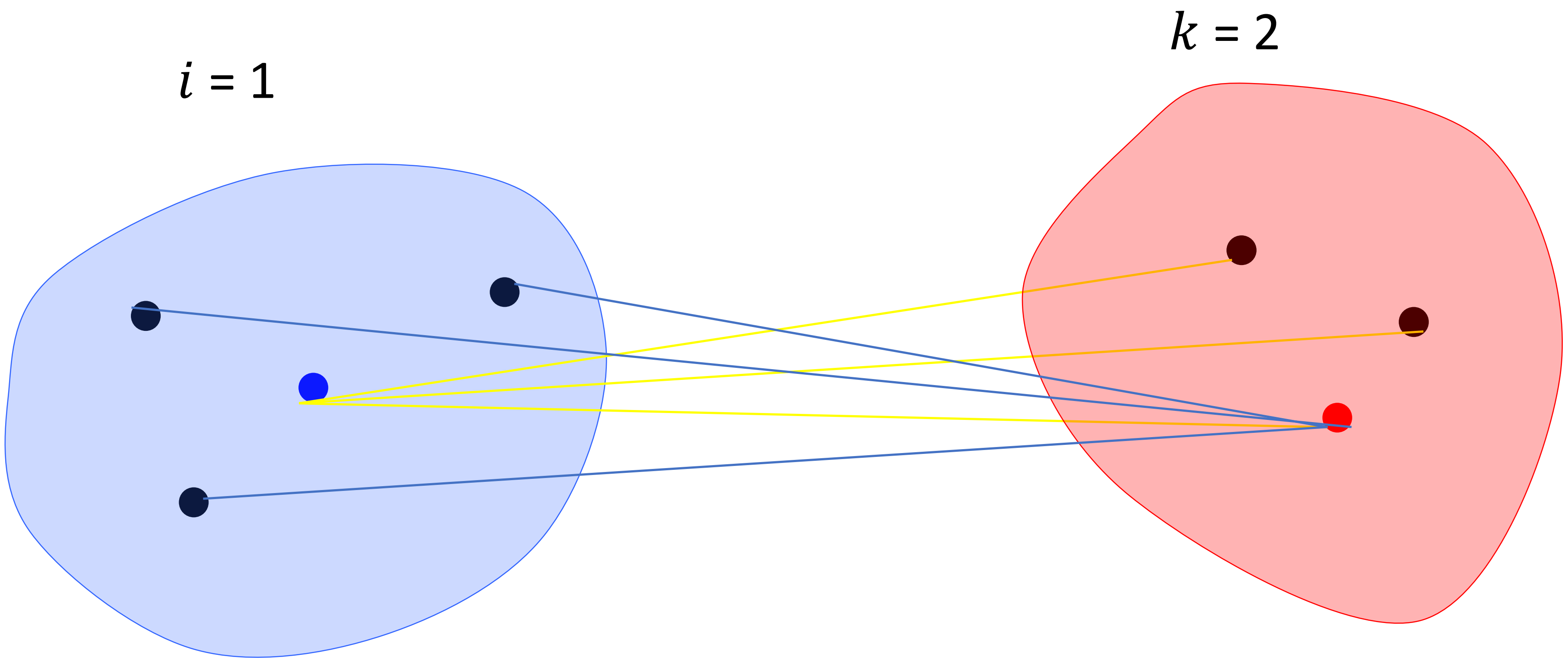
Separation (inter-cluster)

Distance d from a point x to a set C is defined as $:= d(x, C)$



Separation (inter-cluster)

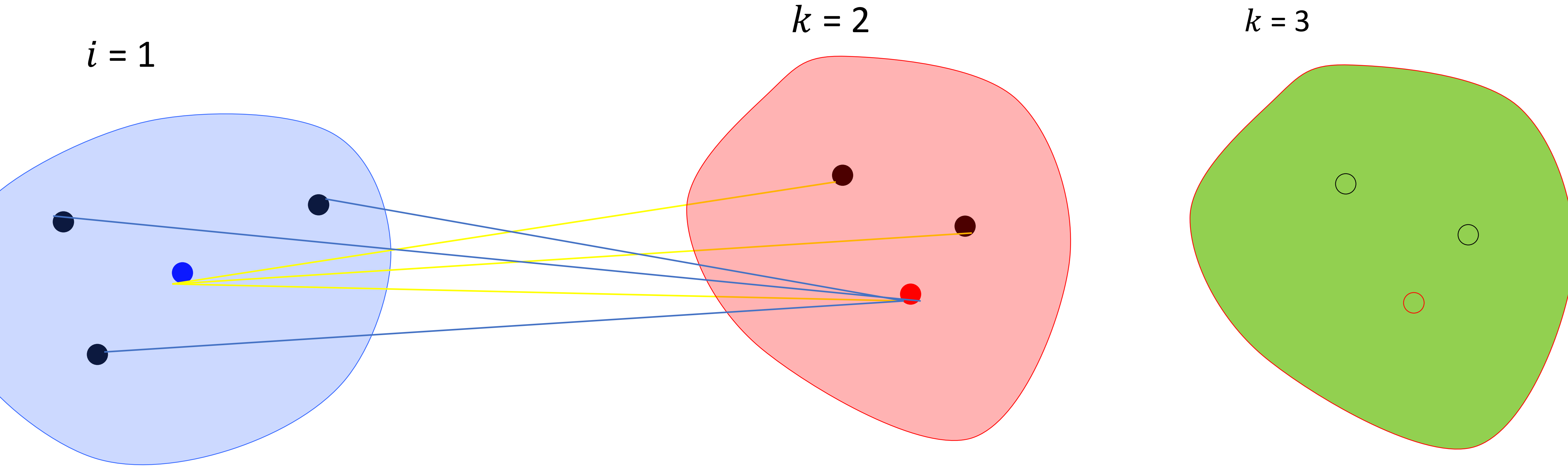
$K = 2$



$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

Separation (inter-cluster)

$K = 3$



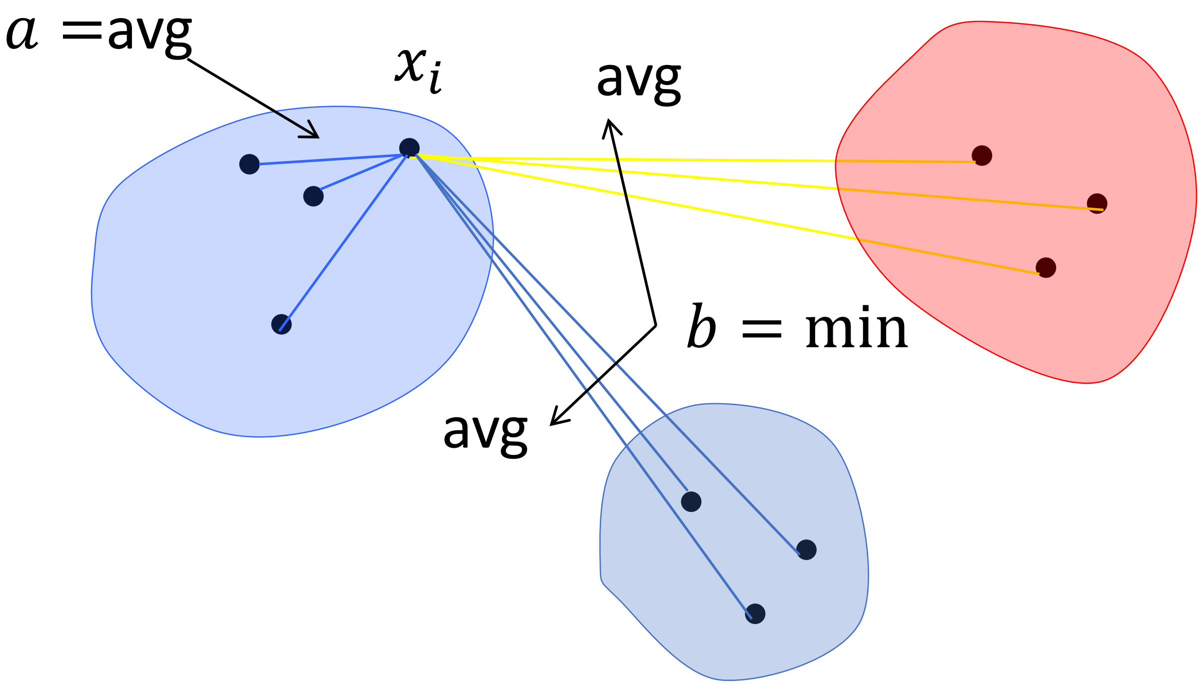
$$b(i) = \min_{k \neq i} \boxed{\frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)}$$

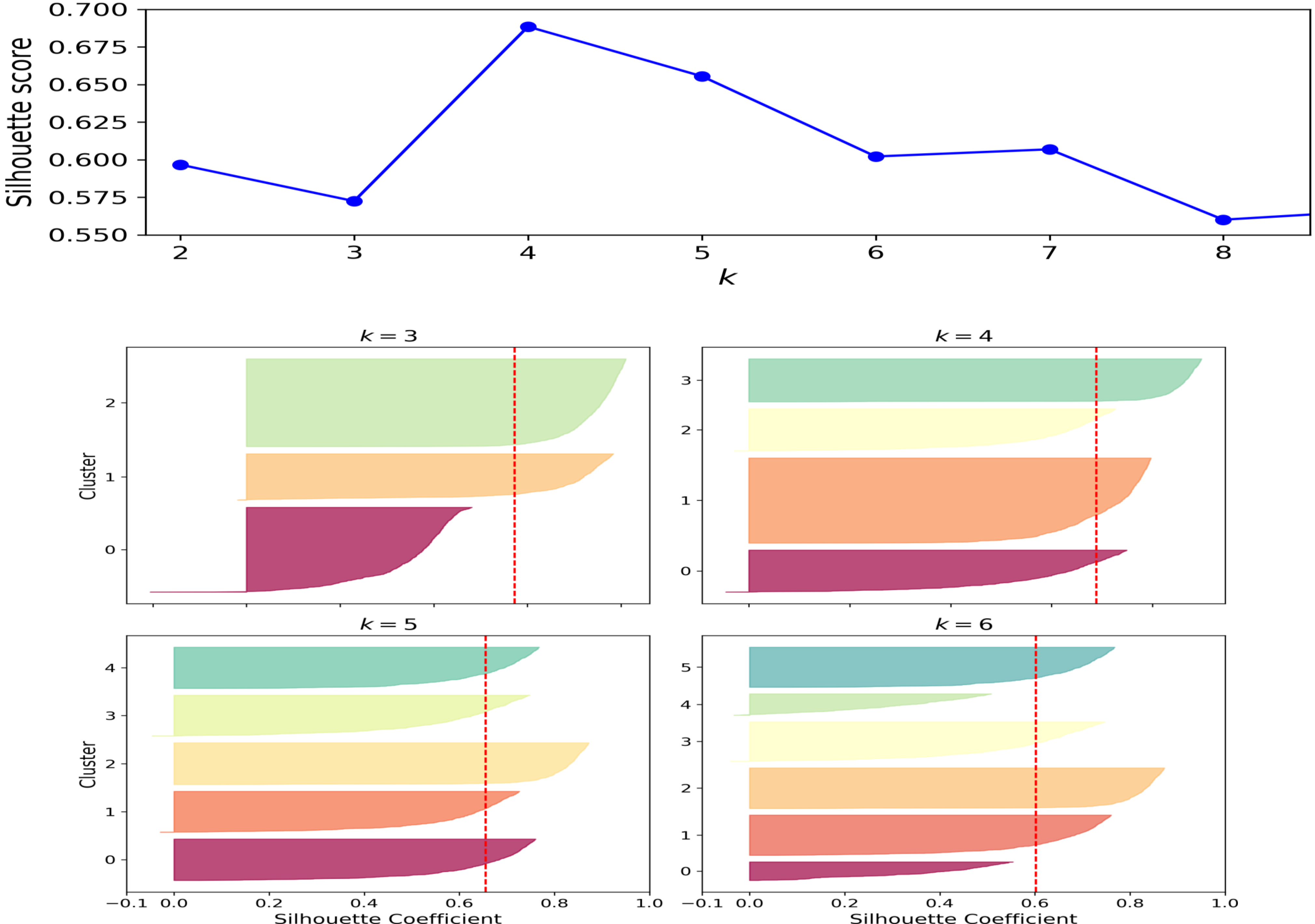
Silhouette Coefficient

- Essentially, combines cohesion and separation into a single number
- Let C_i be cluster of point x_i
  - Let a be average of $d(x_i, y)$ for all y in C_i
  - For $C_i \neq C_j$, let b_j be avg $d(x_i, y)$ for y in C_j
  - Let b be minimum of b_j
- Then, $\boldsymbol{S(x_i) = \frac{b-a}{\max(a,b)}}$
- if $|C_i| > 1$, else $S(x_i) = 0$

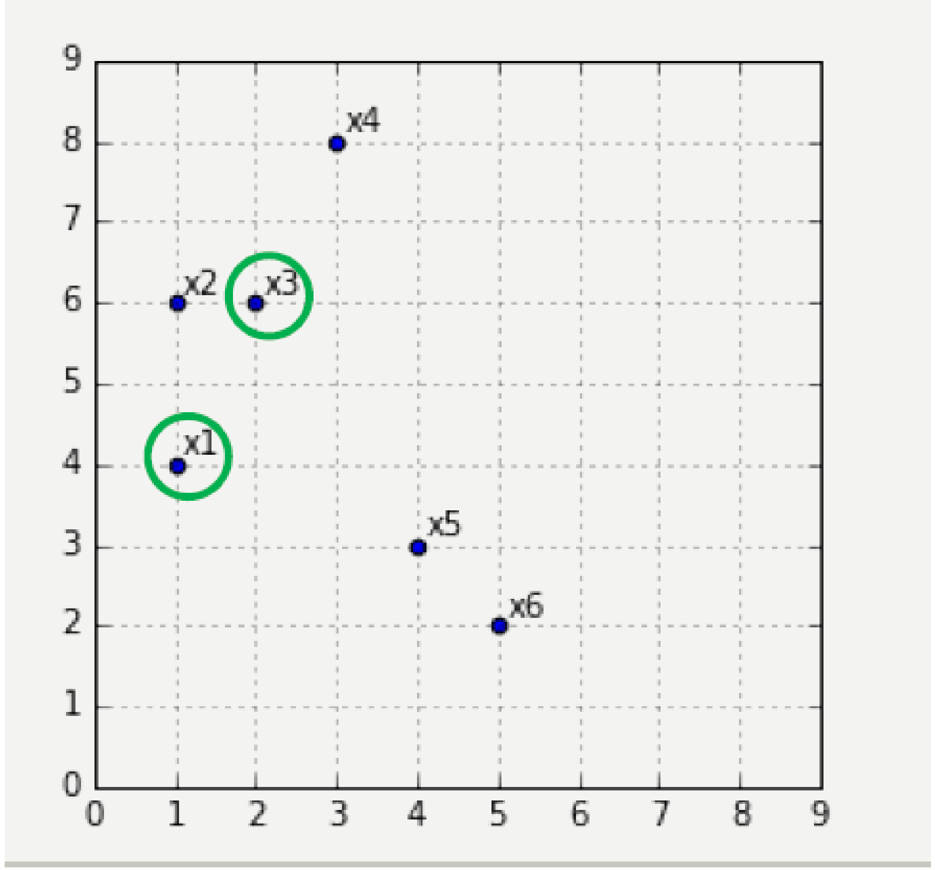
Silhouette Coefficient

- The idea...




Silhouette score
k
k = 3
k = 4
k = 5
k = 6
Cluster
Silhouette Coefficient

- Compute the Silhouette coefficient for



>_ Code

Homework assignment

- Using Scikit learn, find the Silhouette coefficient for of the dataset in the previous homework assignment $K = 2, 3,4,5,6$. Plot the Silhoute coefficient as function of K.