K-means algorithm and related

K. Gibert

Knowledge Engineering and Machine Learning group at

Intelligent Data Science and Artificial Intelligence Research Center

Universitat Politècnica de Catalunya, (IDEAI-UPC) Barcelona

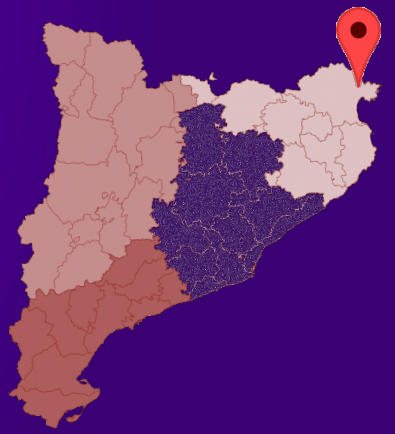
Dean of Illustrious Professional College of Informatics Engineering of Catalonia

President and founder of donesCOEINF

https://www.eio.upc.edu/en/homepages/karina,

karina.gibert@upc.edu, karina.gibert, @karinagibertk

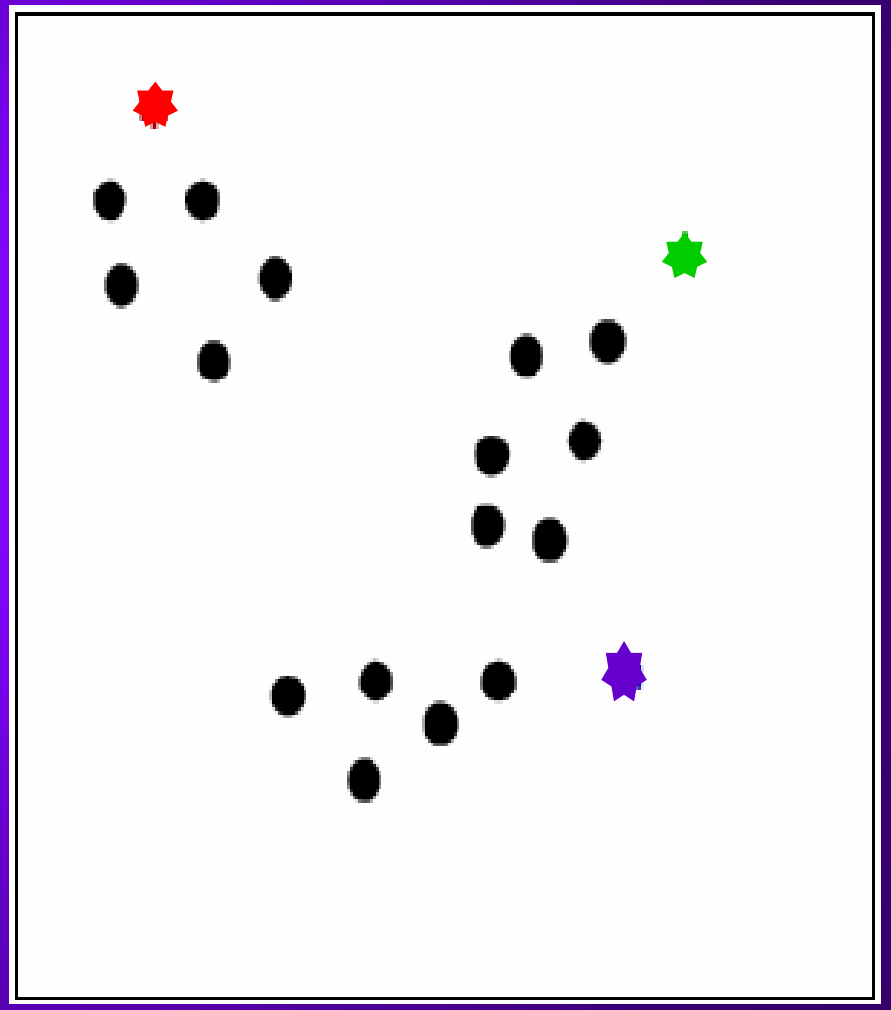
Figuerenca



Abril, 2024

K-Means algorithm *[McQueen 67]*

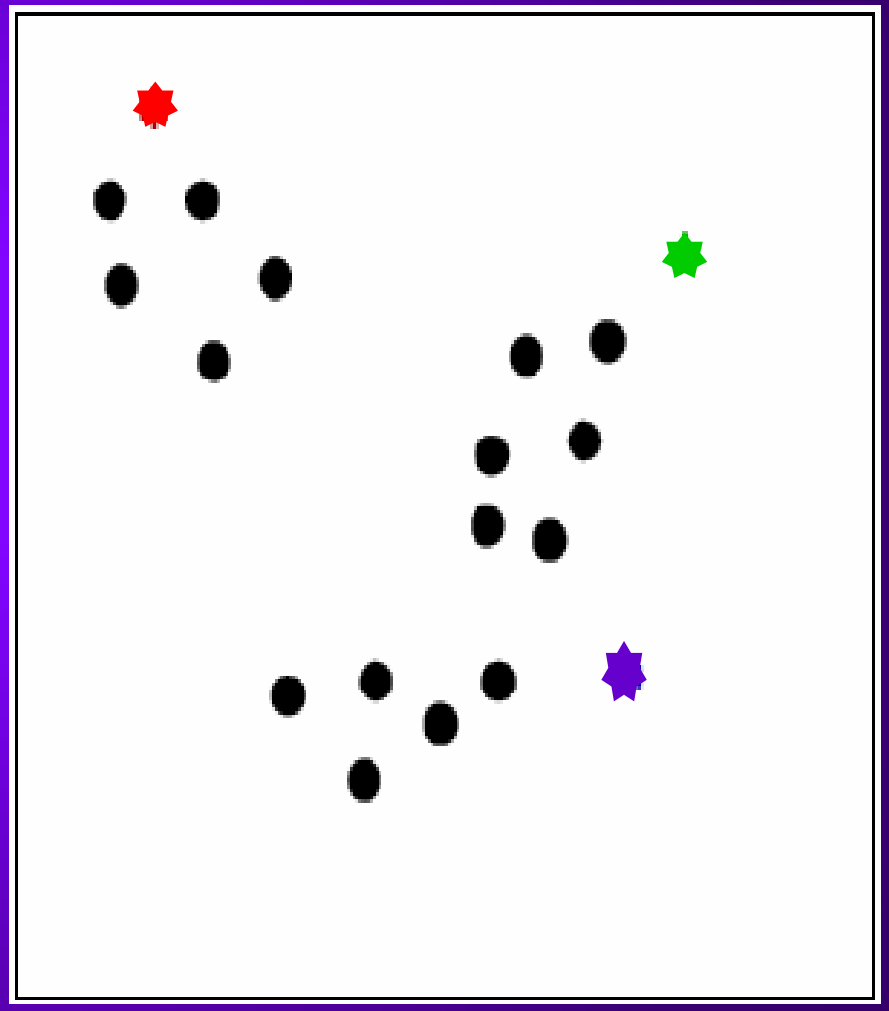
- Select k
- Select k seeds (random?...)
  - *K-Medoids*
    (use Medians as seeds)



MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

K-Means algorithm

- Select k
- Select k seeds (random?...)

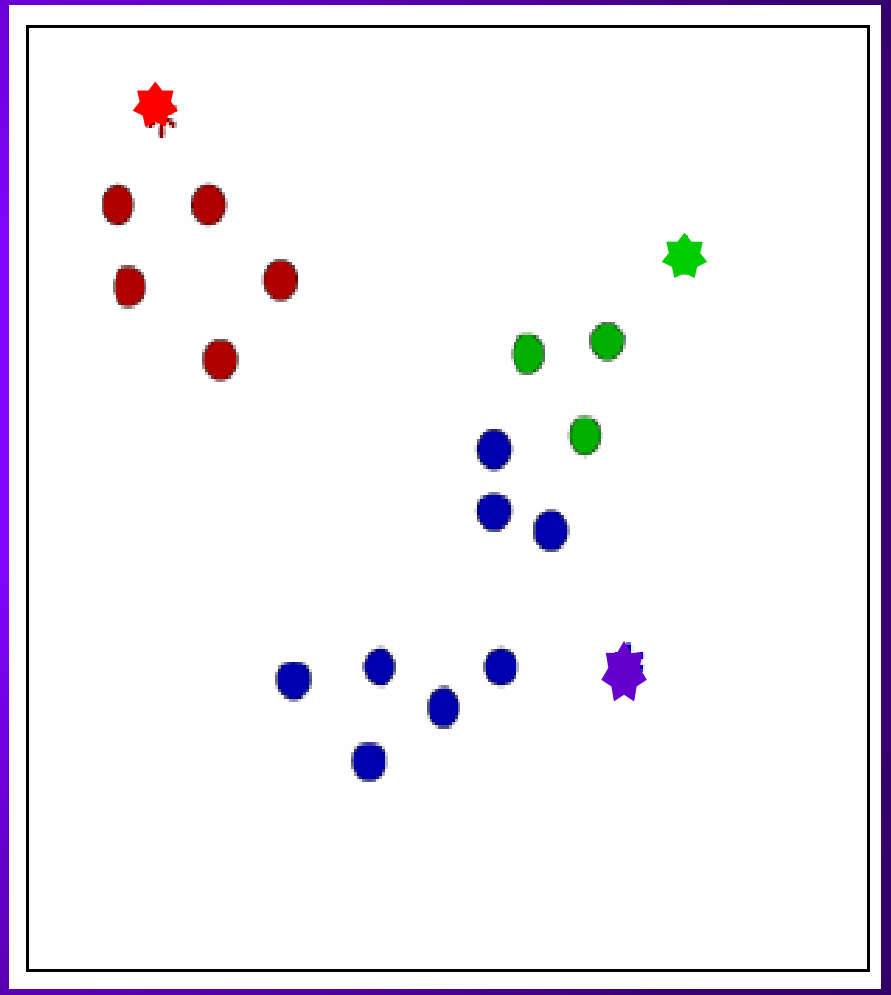


K-Means algorithm

- Select k
- Select k seeds (random?...)

- Assign objects to seeds (class)

$$\min_{\forall g_c} \sum_{\forall c} \sum_{\forall i \in C} ||x_i - gc||^2$$

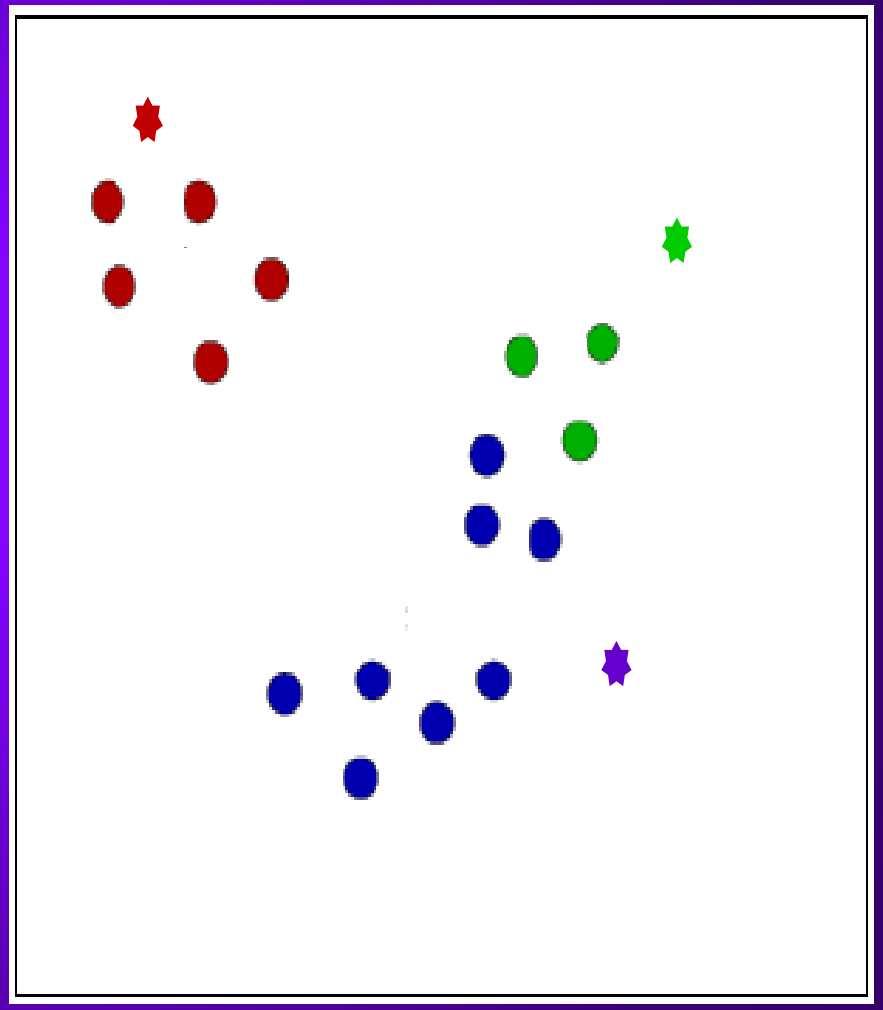


K-Means algorithm

- Select k
- Select k seeds (random?...)

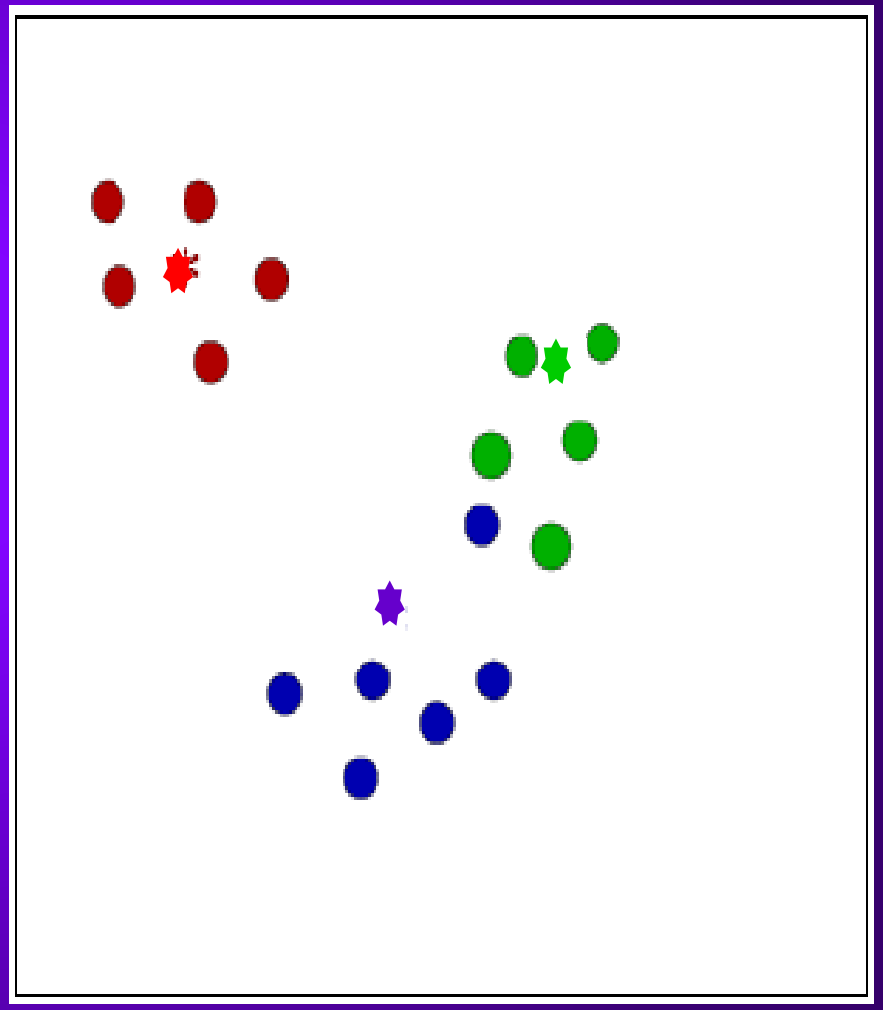
- Assign objects to seeds (class)

- Update seeds accordingly



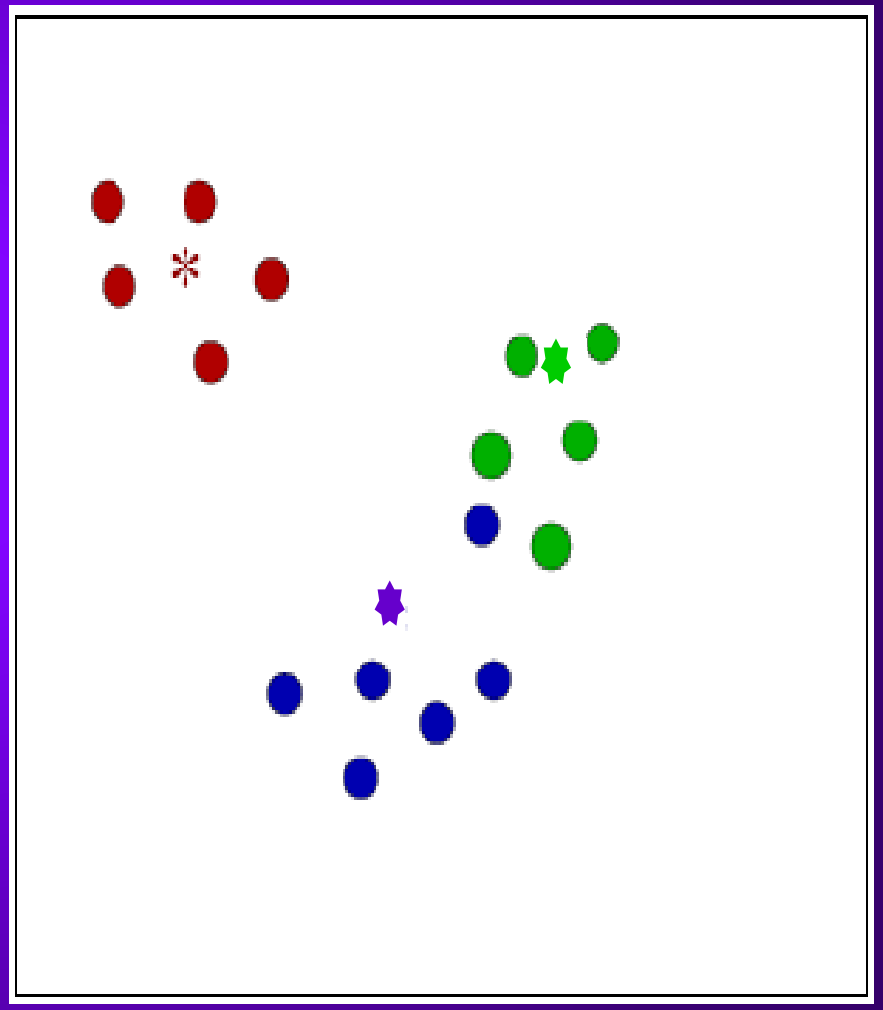
K-Means algorithm

- Select k
- Select k seeds (random?...)
- Assign objects to seeds (class)
- Update seeds accordingly
- Iterate till no changes found



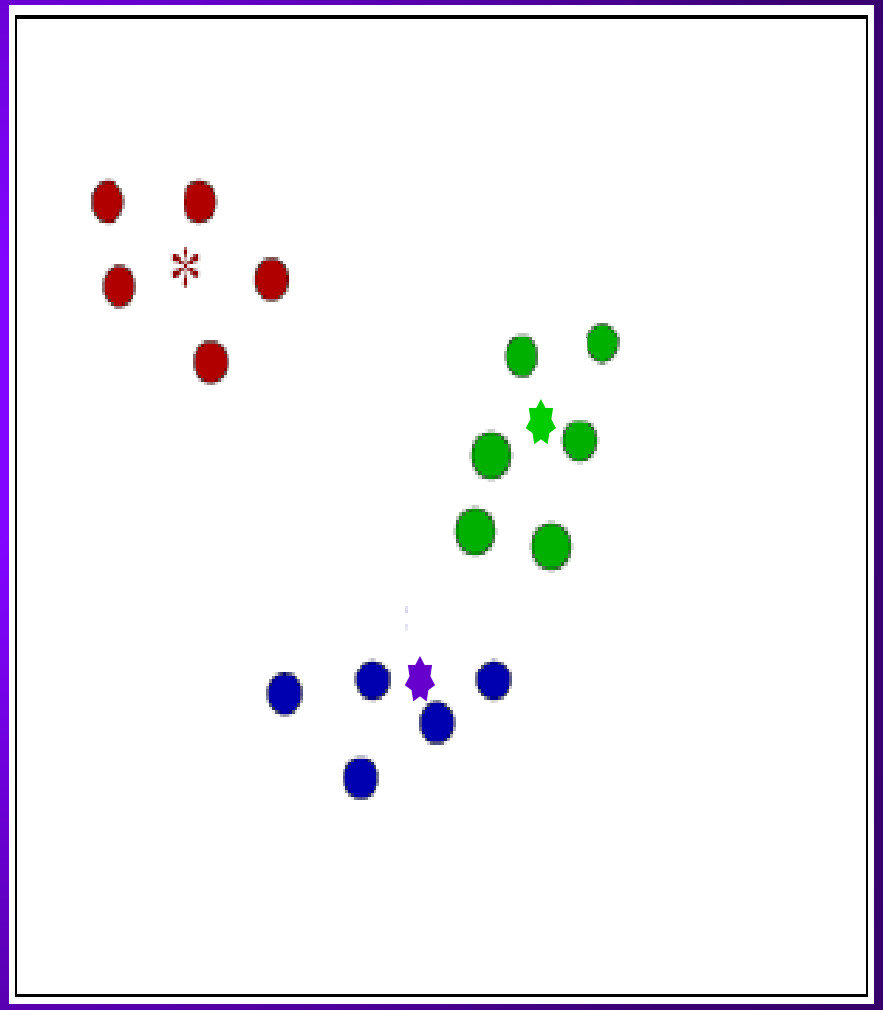
K-Means algorithm

- Select k
- Select k seeds (random?...)
- Assign objects to seeds (class)
- Update seeds accordingly
- Iterate till no changes found



K-Means algorithm

- Select k
- Select k seeds (random?...)
- Assign objects to seeds (class)
- Update seeds accordingly
- Iterate till no changes found



K-Means algorithm

- Drawbacks

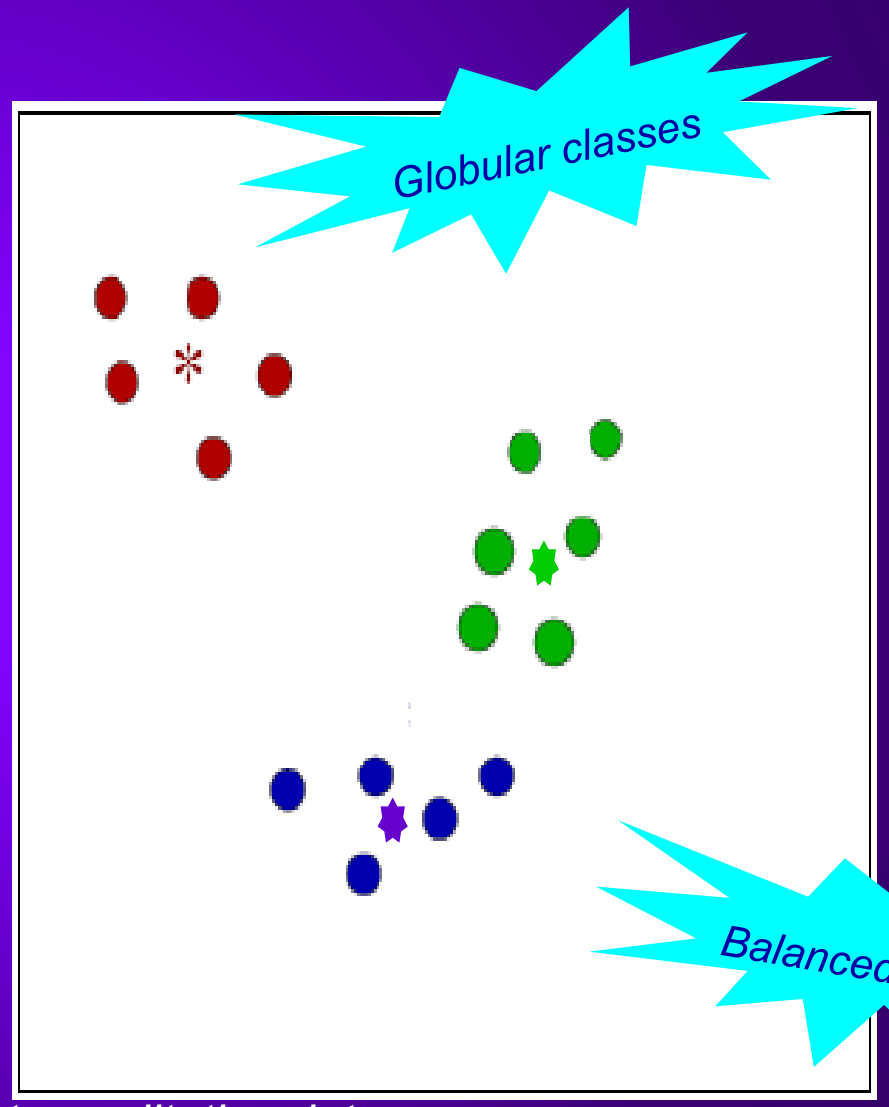
- Order depending
  - Local optima

- Sensitive to outliers

- SSE decreases with k
  no goodness indicator

- Fast, but not scalable
  Fast K-Means (Triangular inequality)

- K is an input
  (X-Means: kd-trees and
  Bayesian information criterion)

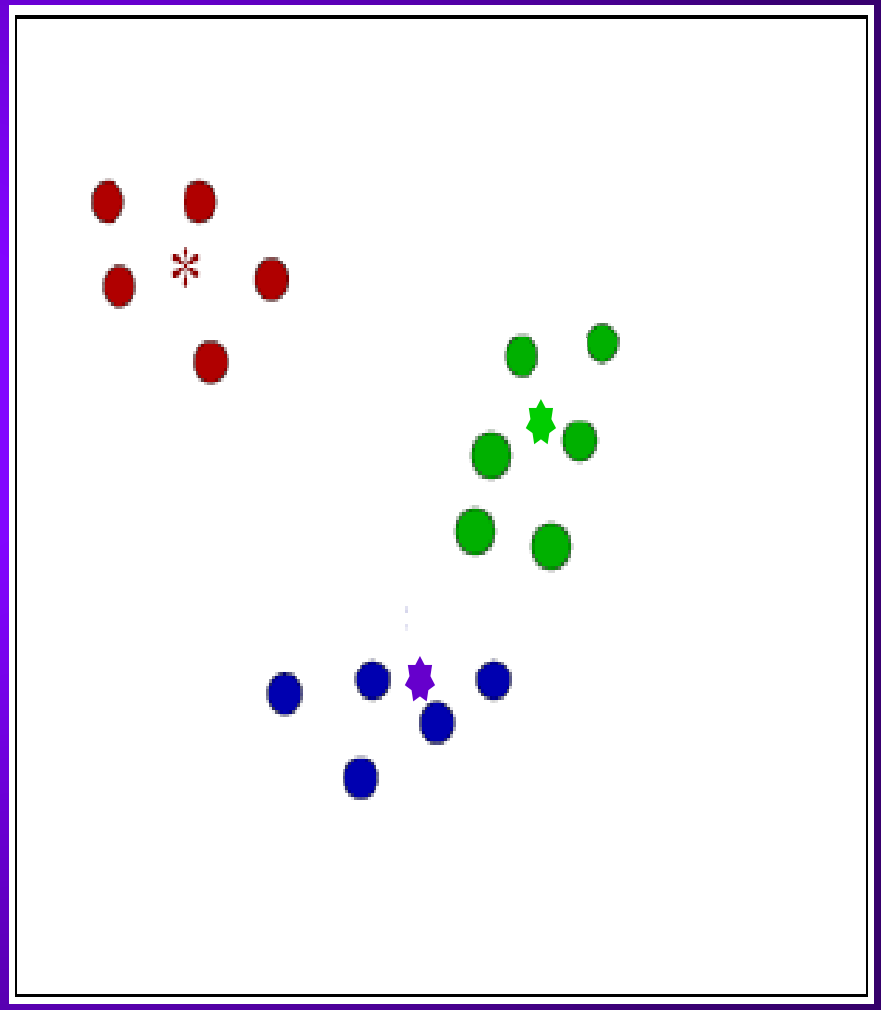


Non directly aplicable to qualitative data

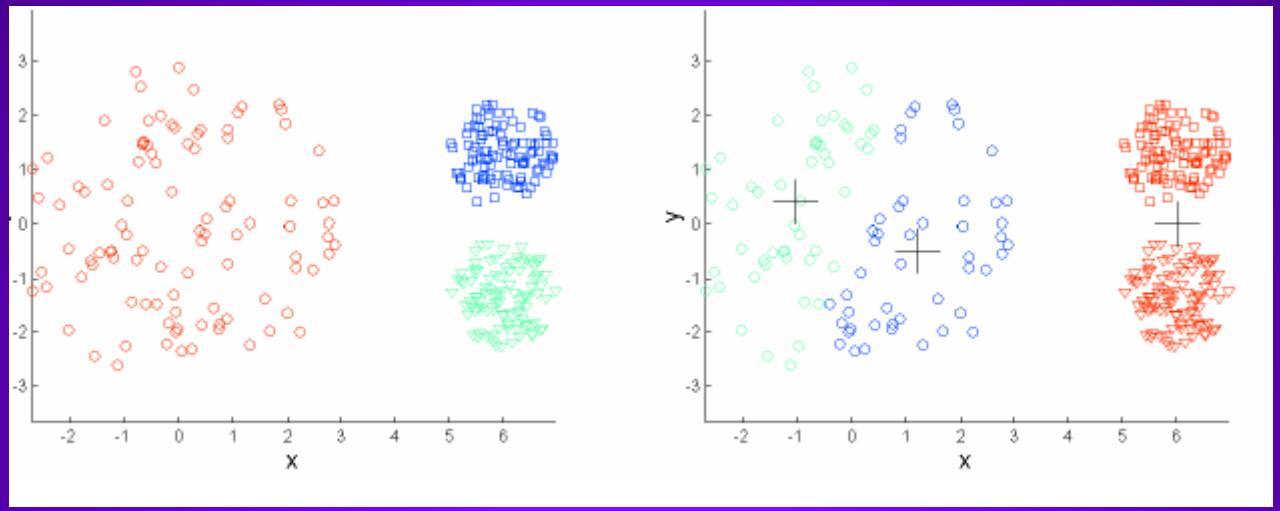
K-Means algorithm

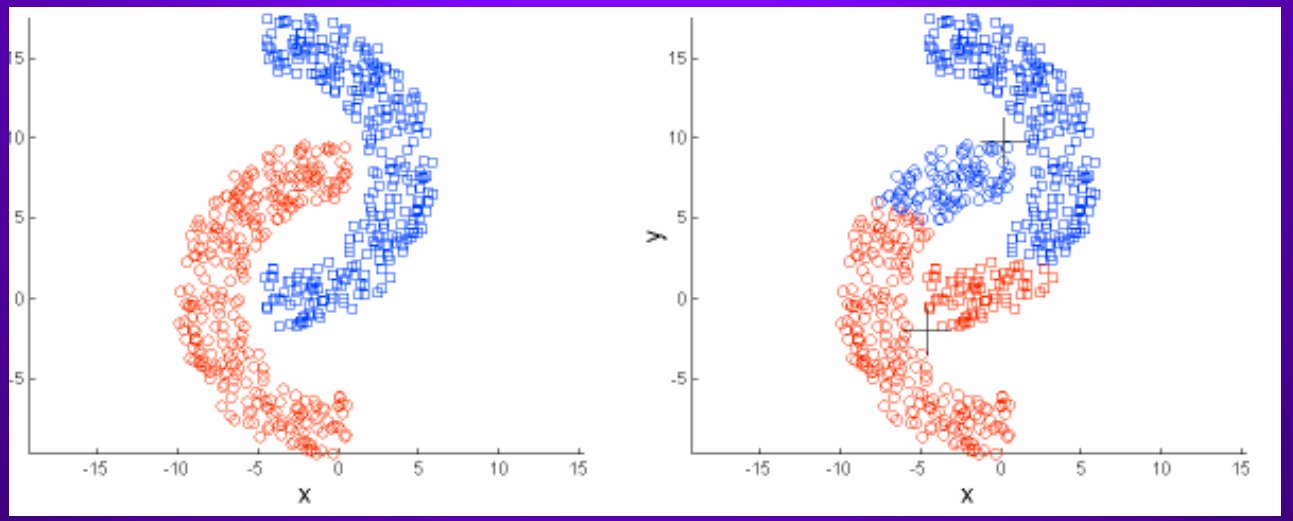
Technical assumptions

- Globular classes
- Balanced data
- No-outliers



K-Means algorithm





Case Study: Fisher's Iris Flower

Iris setosa

iris.csv ***

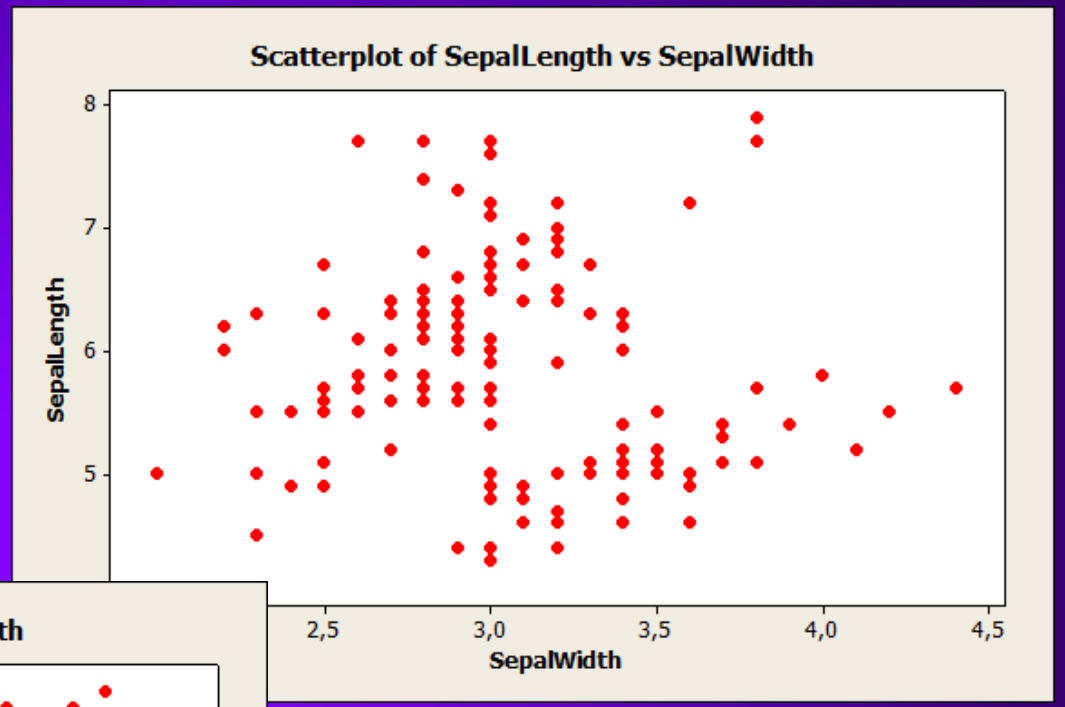
↓	C1	C2	C3	C4	C5-T
	SepalLength	SepalWidth	PetalLength	PetalWidth	Species
1	5,1	3,5	1,4	0,2	setosa
2	4,9	3,0	1,4	0,2	setosa
3	4,7	3,2	1,3	0,2	setosa
4	4,6	3,1	1,5	0,2	setosa
5	5,0	3,6	1,4	0,2	setosa
6	5,4	3,9	1,7	0,4	setosa
7	4,6	3,4	1,4	0,3	setosa
8	5,0	3,4	1,5	0,2	setosa
9	4,4	2,9	1,4	0,2	setosa
10	4,9	3,1	1,5	0,1	setosa
11	5,4	3,7	1,5	0,2	setosa
12	4,8	3,4	1,6	0,2	setosa
13	4,8	3,0	1,4	0,1	setosa
14	4,3	3,0	1,1	0,1	setosa
15	5,8	4,0	1,2	0,2	setosa
16	5,7	4,4	1,5	0,4	setosa
17	5,4	3,9	1,3	0,4	setosa

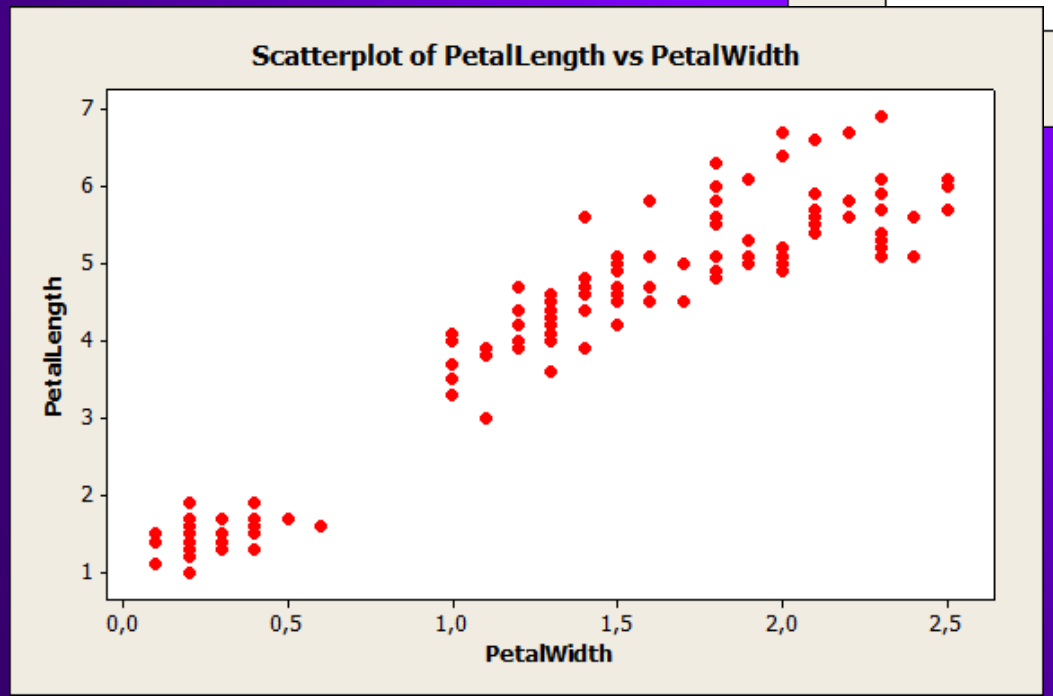
Iris versicolor

Fisher 1936
Anderson measured
150 flowers

Iris virginica

Case Study: Fisher's Iris Flower



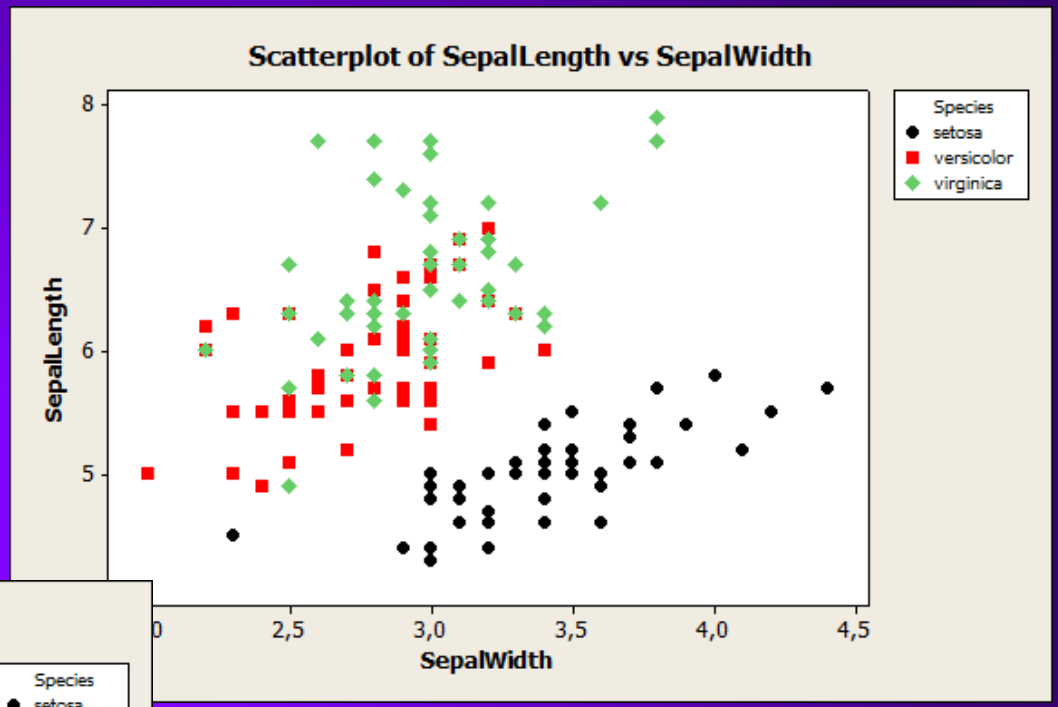


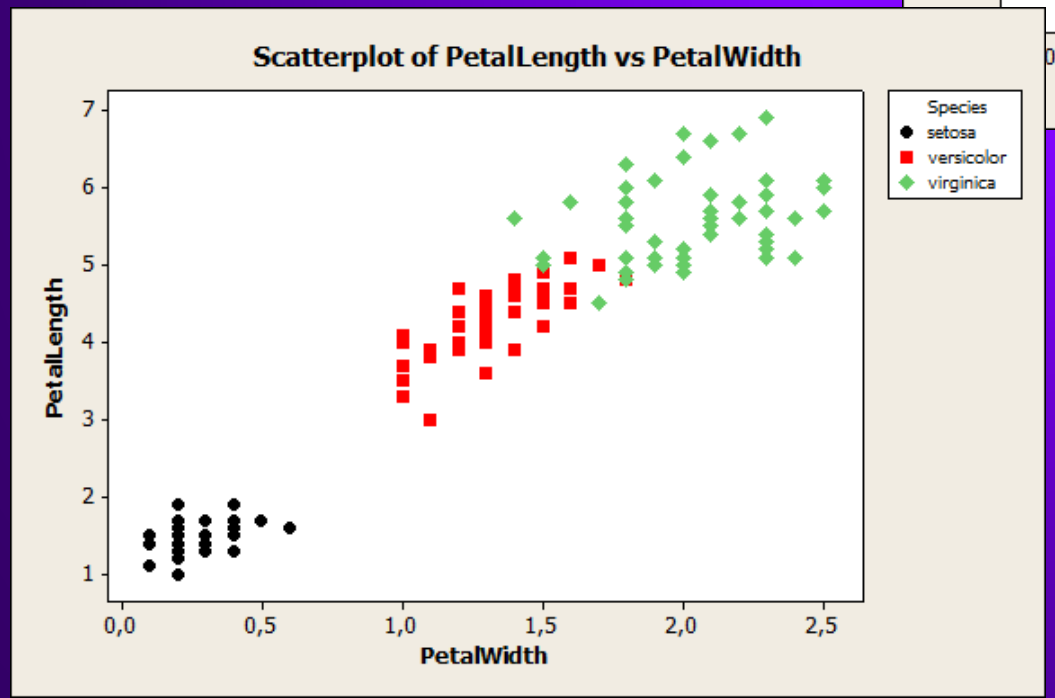
Case Study: Fisher's Iris Flower

Iris setosa

Iris versicolor

Iris virginica





Feature Weighting

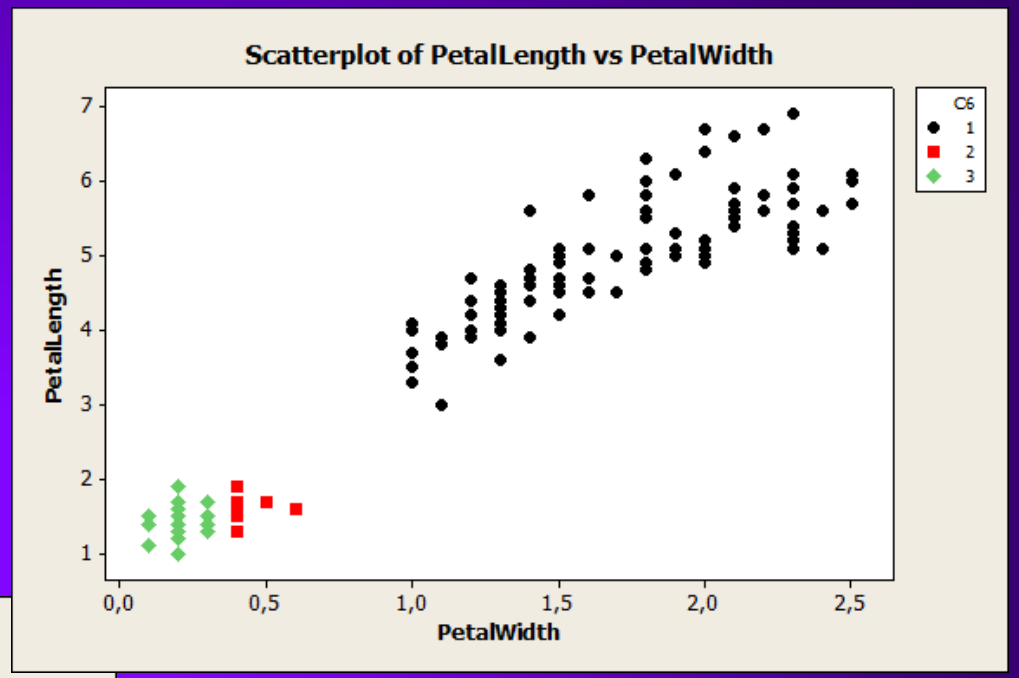
PL: 9,08
PW: 9,7
SL: 0.3
SW: 0.05

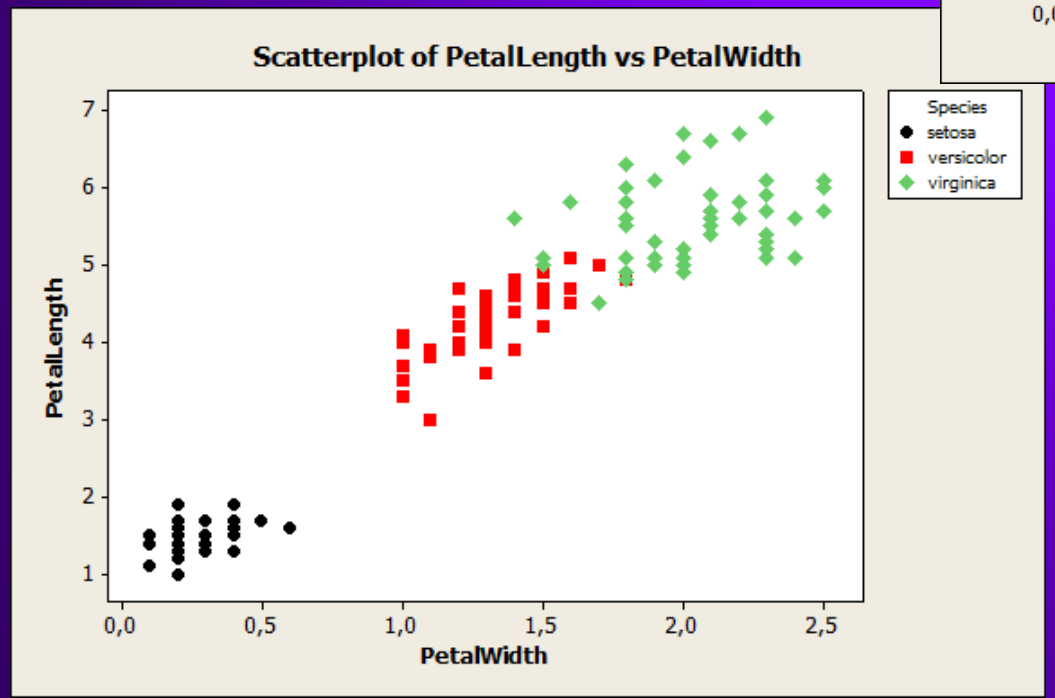
Case Study: Fisher's Iris Flower


Iris setosa


Iris versicolor

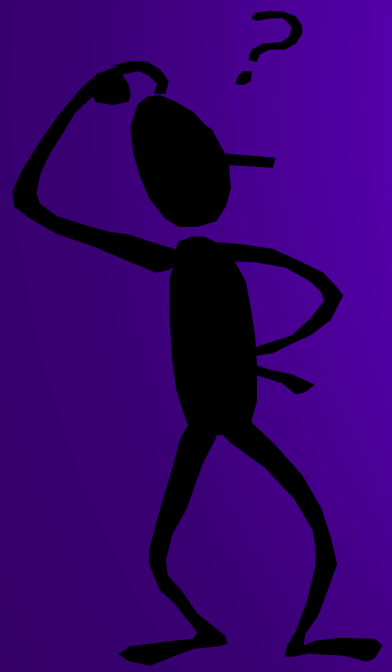

Iris virginica





3-kMeans

Karina Gibert
Dpt. Statistics and Operation Research
Knowledge Engineering and Machine Learning Research group at Intelligent Data Science and Artificial Intelligence Research Center
Research Institute of Sience and Technology of Sustainability
Universitat Politècnica de Catalunya-BarcelonaTech (Spain)
karina.gibert@upc.edu
www.eio.upc.edu/en/homepages/karina



Are there any questions?...