Clustering for Different Scales of Measurement - the Gap-Ratio Weighted K-means Algorithm

Joris Guérin **Olivier Gibaru**
Eric Nyiri **Stéphane Thiery**

Arts et Métiers ParisTech

AIAP 2017 - May 27, 2017

Introduction

Preliminaries

Gap-Ratio K-means

Experimental Validation

Conclusion

Outline

Introduction

Preliminaries

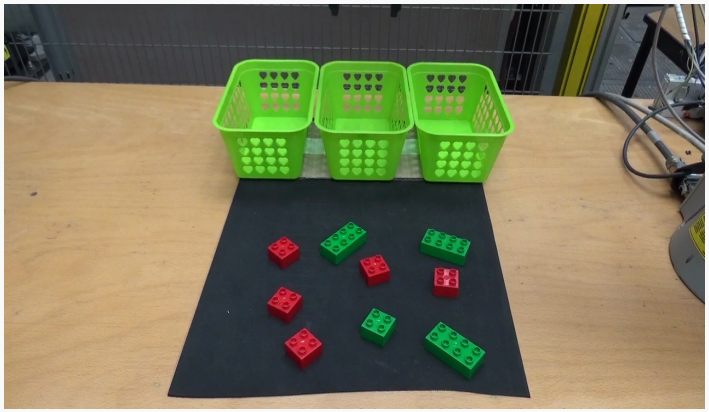
Gap-Ratio K-means

Experimental Validation

Conclusion

Motivations

Objective



Cluster Lego bricks based on shape and color features.

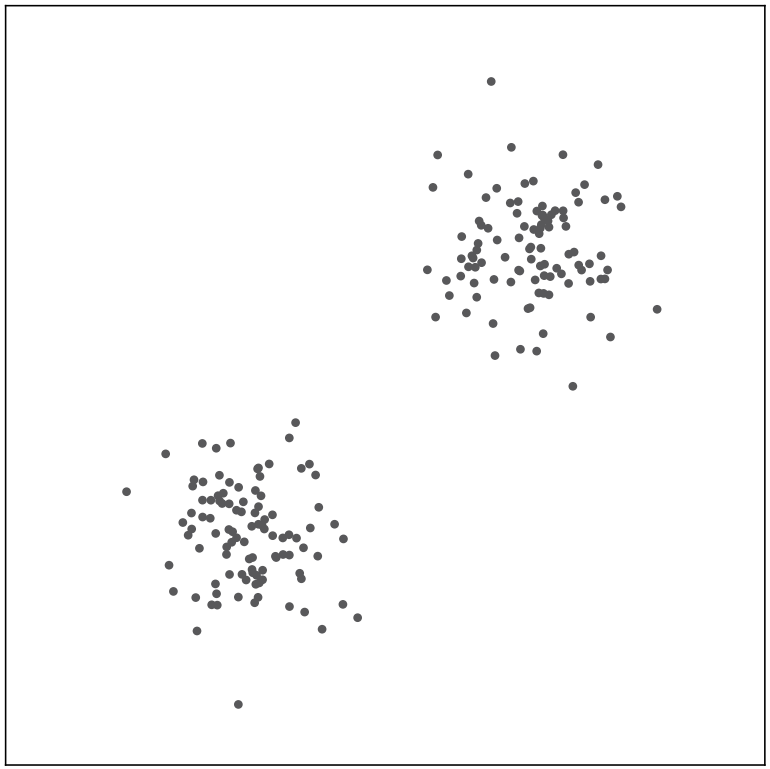
Challenges

- Widely spread data (Unmastered lighting conditions)
- Interval-scale features (RGB)

robot demo video

The clustering problem

Before Clustering



After Clustering

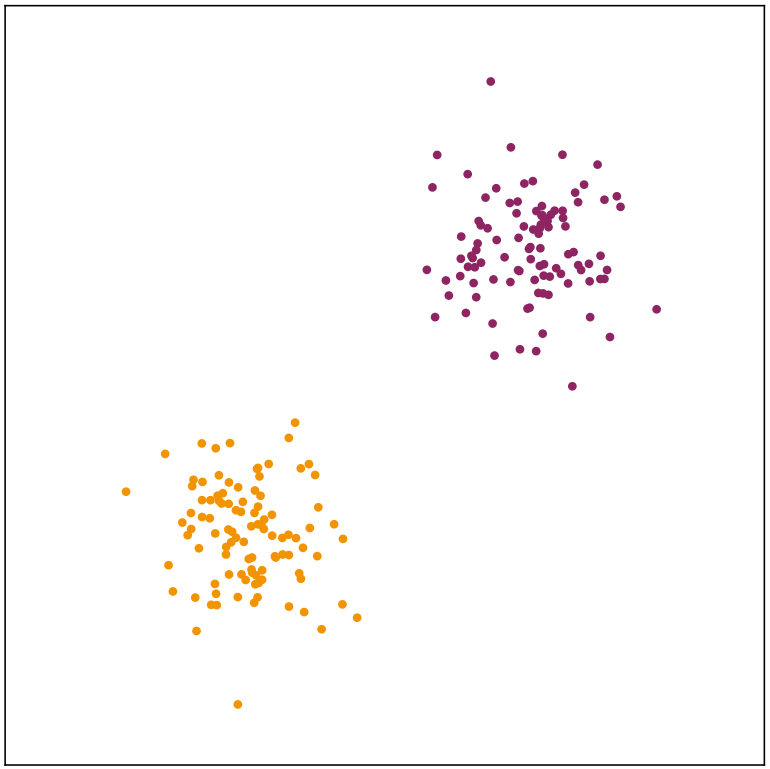


Illustration of the clustering problem in 2D

Outline

Introduction

Preliminaries

Gap-Ratio K-means

Experimental Validation

Conclusion

K-means clustering

Formulation for M data points and K desired clusters.

$$\underset{A,\, c}{\text{Minimize}} \quad \sum_{i=1}^{M}\sum_{k=1}^{K} a_{ik} \times d(x_i, c_k),$$

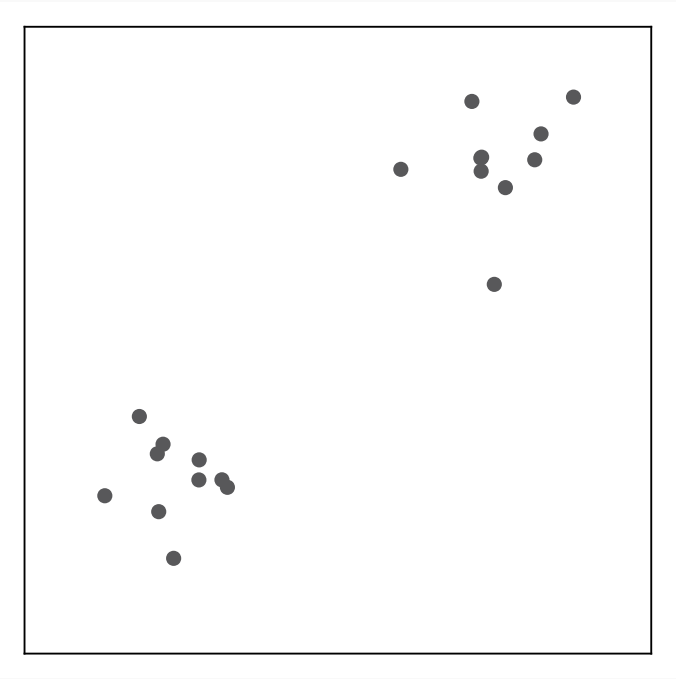
$$\text{subject to} \quad \sum_{k=1}^{K} a_{ik} = 1,\ \forall i \in \{1, ..., M\},$$

$$a_{ik} \in \{0, 1\},\ \forall i, \forall k.$$

With

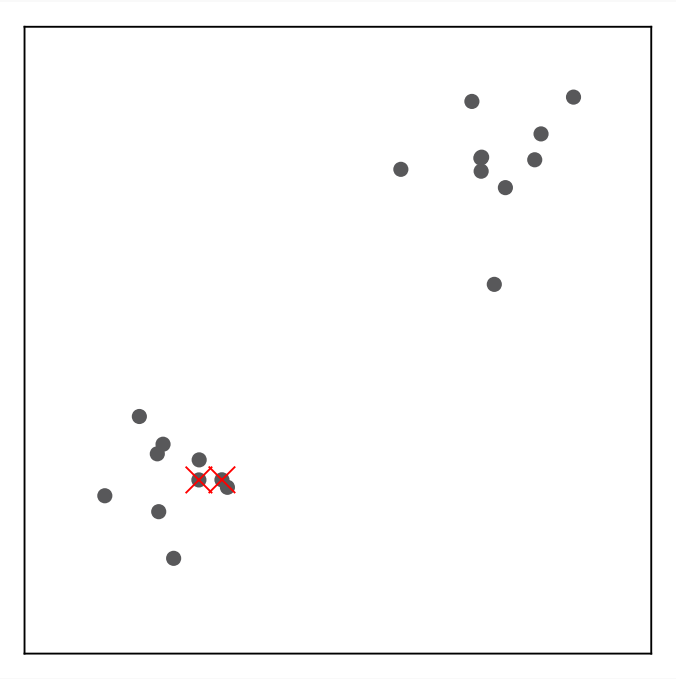
- c_k cluster centers,
- a_{ik} membership binary variables,
- $d(.,.)$ distance metric used.

K-means resolution using Alternating Optimization



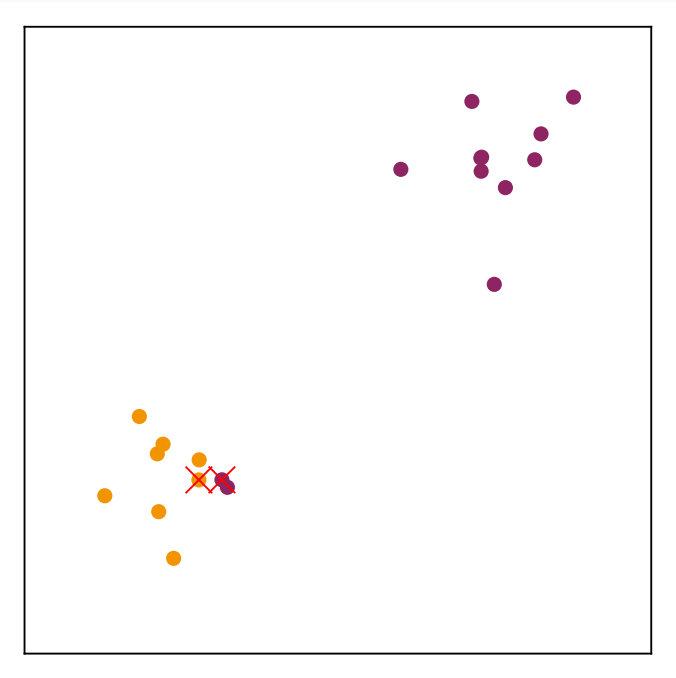
Initial data

K-means resolution using Alternating Optimization



Centroids initialization

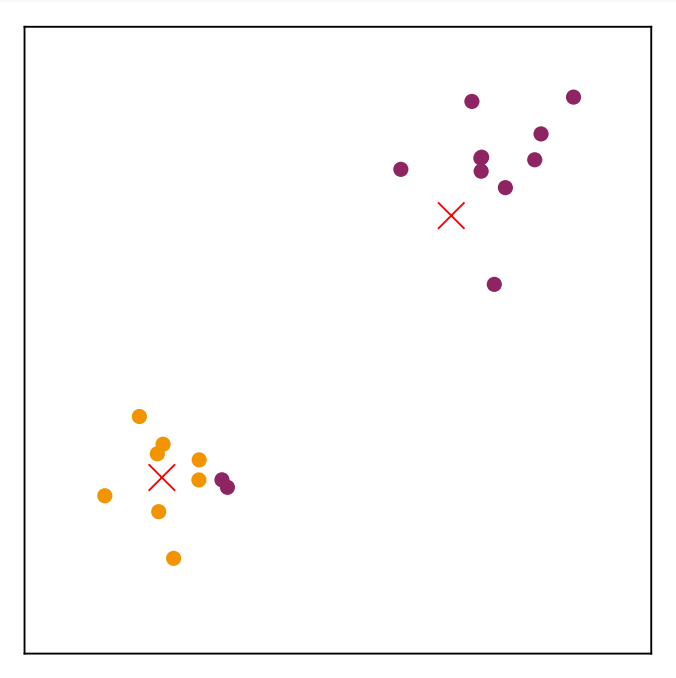
K-means resolution using Alternating Optimization



Classes actualization

$$x_i \in C_l \iff d(x_i, c_l) \leq d(x_i, c_k),\ \forall k \in \{1, ..., K\}.$$

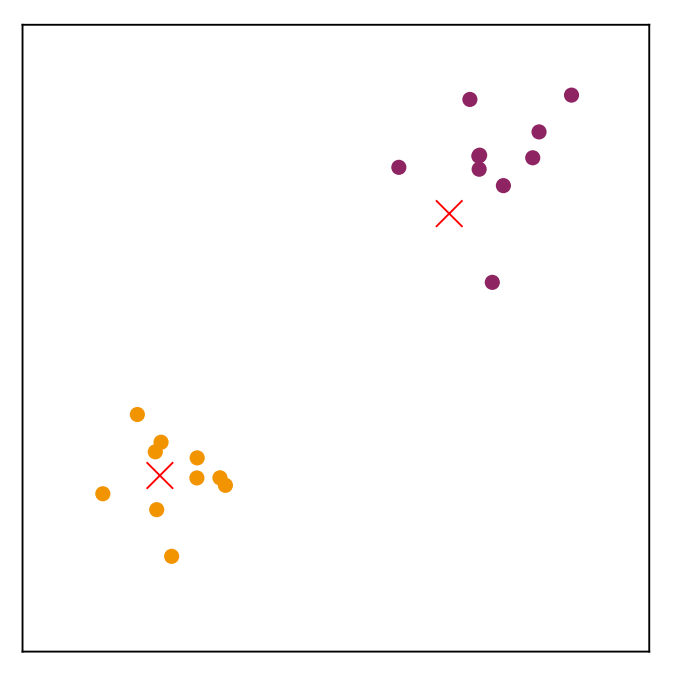
K-means resolution using Alternating Optimization



Centroids update

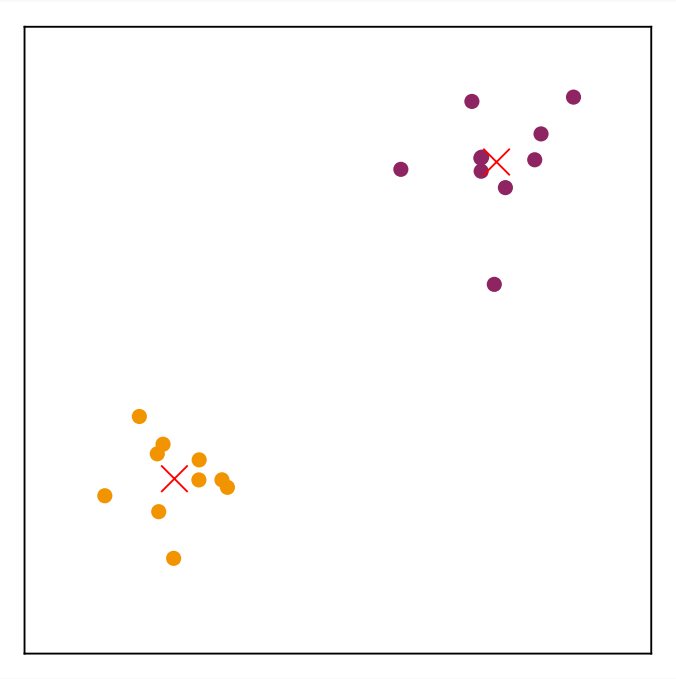
$$c_k = \frac{1}{\sum_{i=1}^{M} a_{ik}} \sum_{i=1}^{M} a_{ik} \times x_i$$

K-means resolution using Alternating Optimization



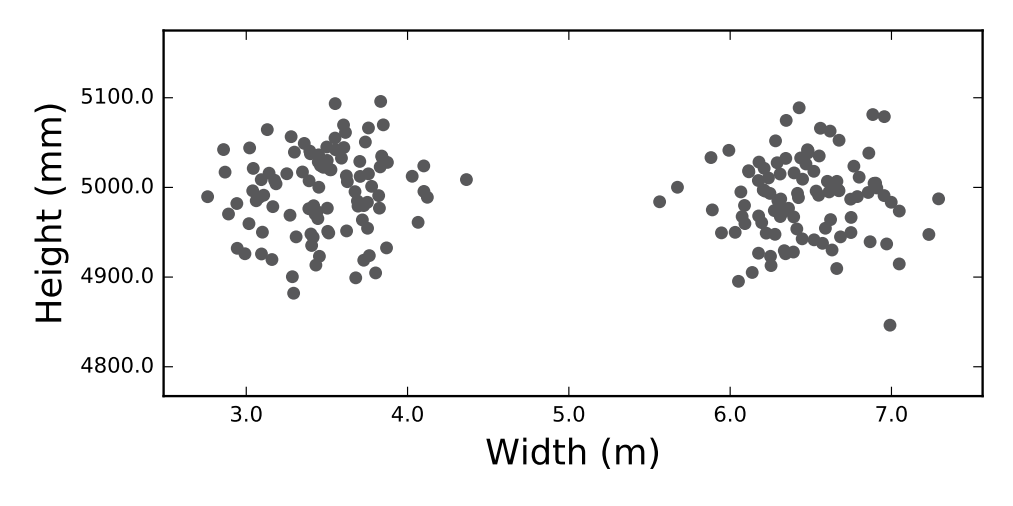
Classes actualization

K-means resolution using Alternating Optimization



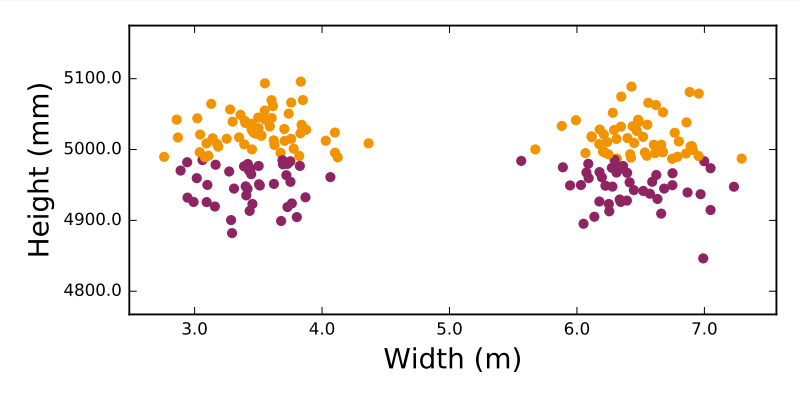
Centroids update

Data normalization

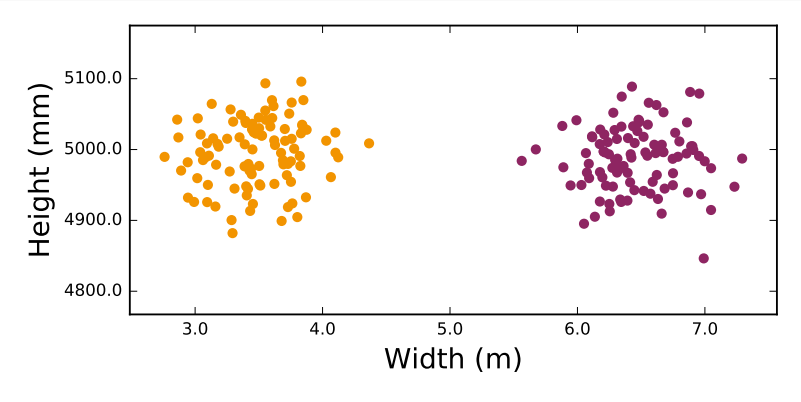


Initial data

Data normalization

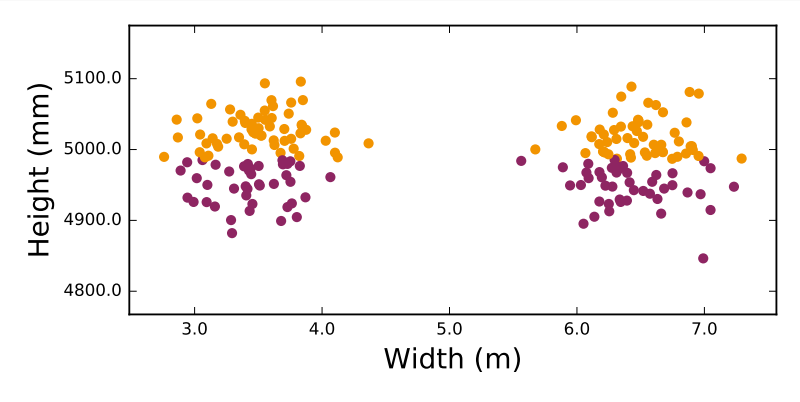


Without data normalization

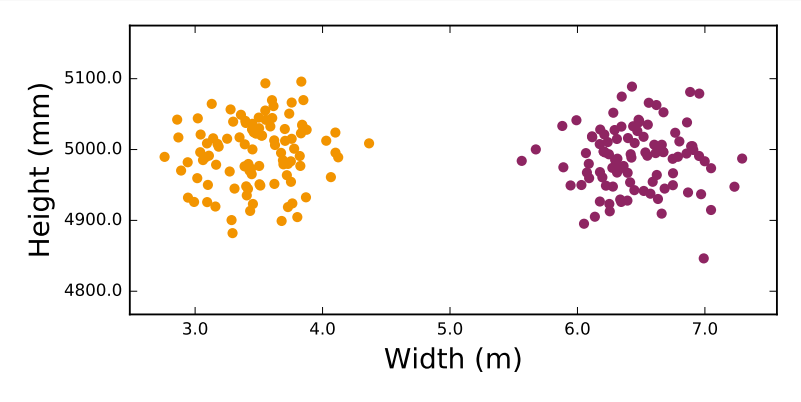


With data normalization

Data normalization



Without data normalization



With data normalization

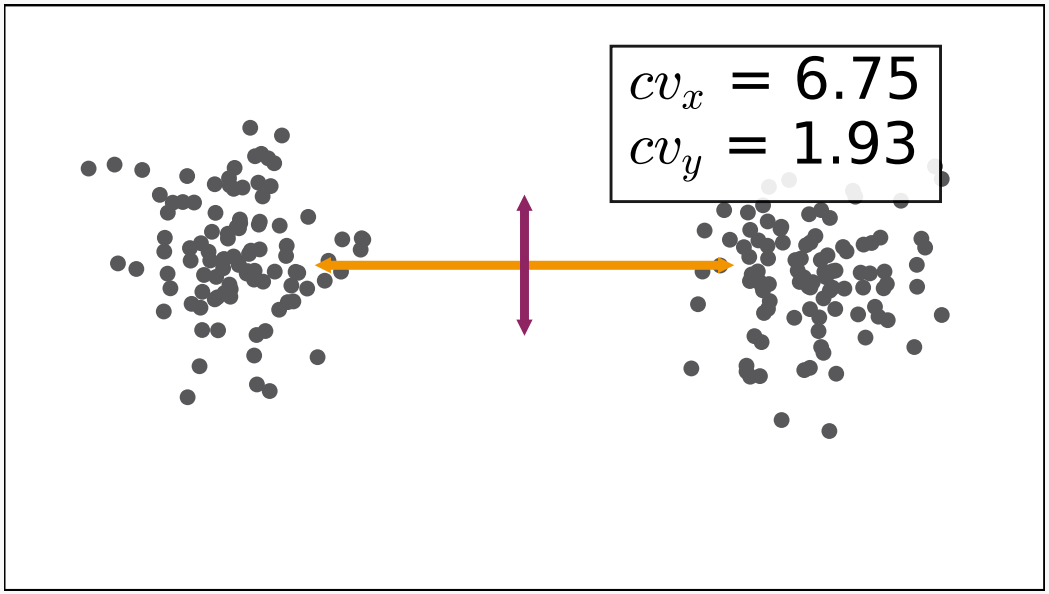
- Issue : With noisy data, loss of valuable information

Weighted K-means

- Store information in weights
- Modify the distance metric using these weights:

$$d(x_i, c_k) = \sqrt{\sum_{j=1}^{N} w_j (x_{ij} - c_{kj})^2}$$

CV K-means

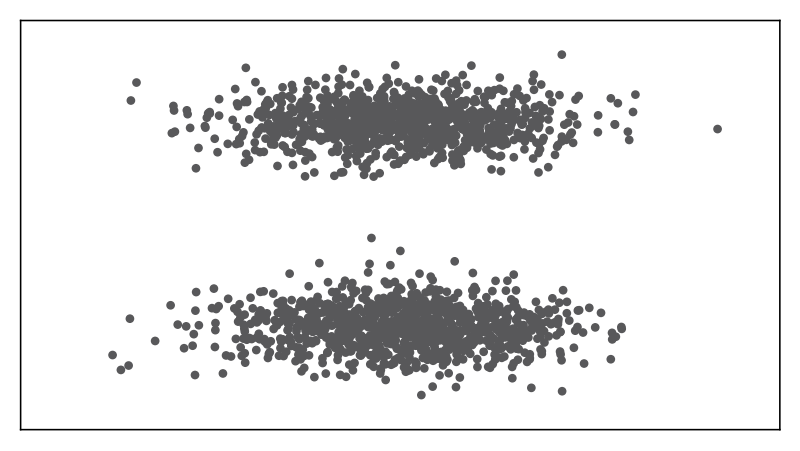


Coefficients of Variation along x and y

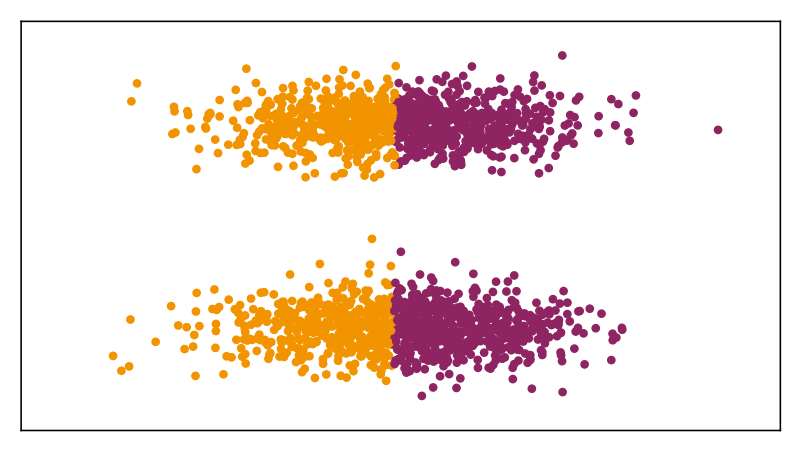
$$cv = \frac{\sigma}{\mu}$$

Issues with CV K-means

- Variance does not always correspond to "natural cluster axis"



Data with high x variance



CV K-means results

Issues with CV K-means

- Issue with interval scale variables

$$cv = \frac{\sigma}{\mu}$$

Outline

Introduction

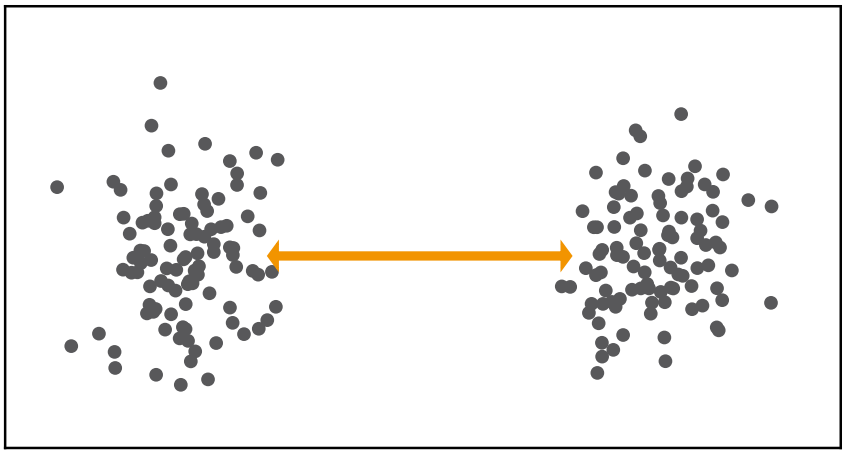
Preliminaries

Gap-Ratio K-means

Experimental Validation

Conclusion

Intuition



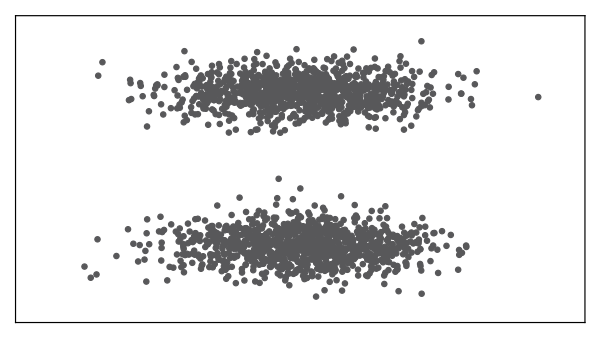
Gap-ratio illustration

$$gr_j = \frac{G_j}{\mu g_j}$$

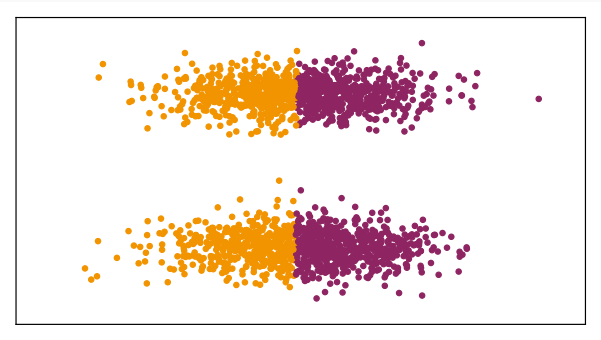
avec

- gr_j gap ratio along dimension j,
- G_j, Largest "gap" along j,
- μg_j Average "gap" along j.

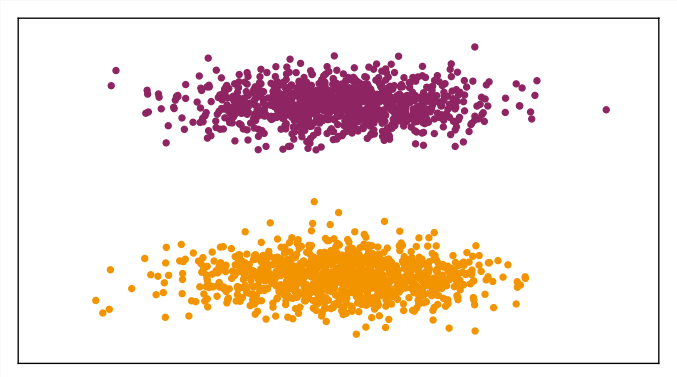
Result on toy problem



Data with high x variance



CV K-means results



GR K-means results

Outline

Introduction

Preliminaries

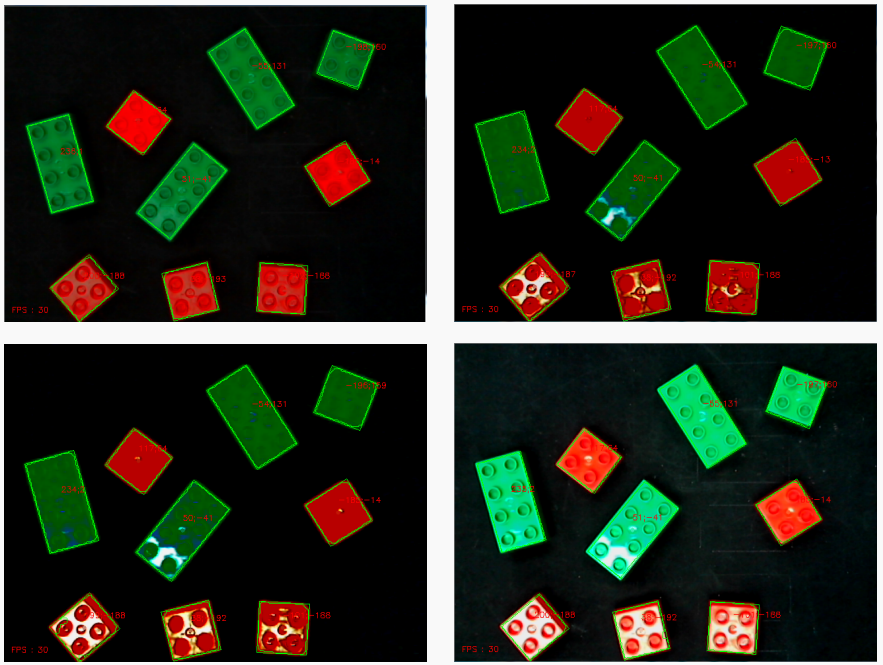
Gap-Ratio K-means

Experimental Validation

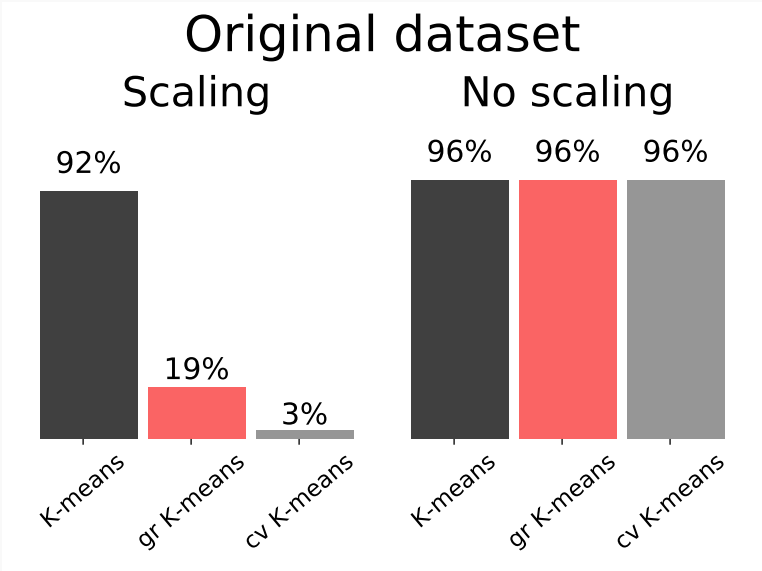
Conclusion

Experiment description

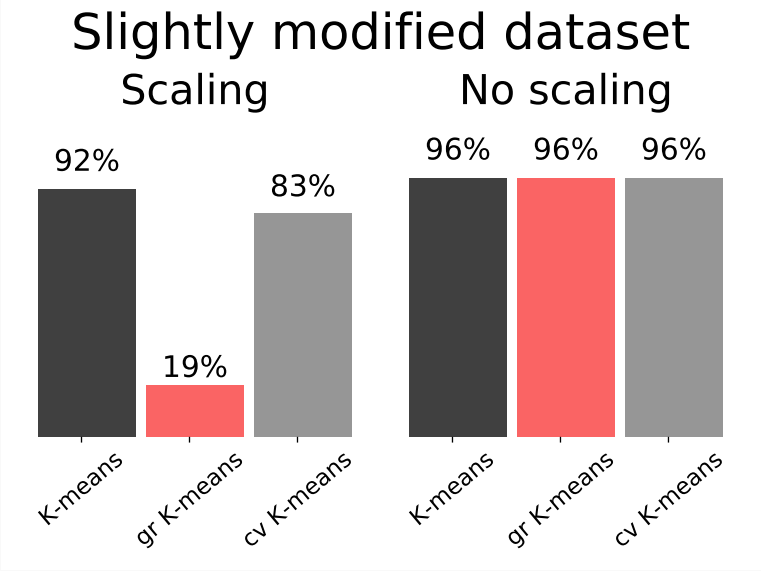
- 98 Different position and lighting configurations.



Results



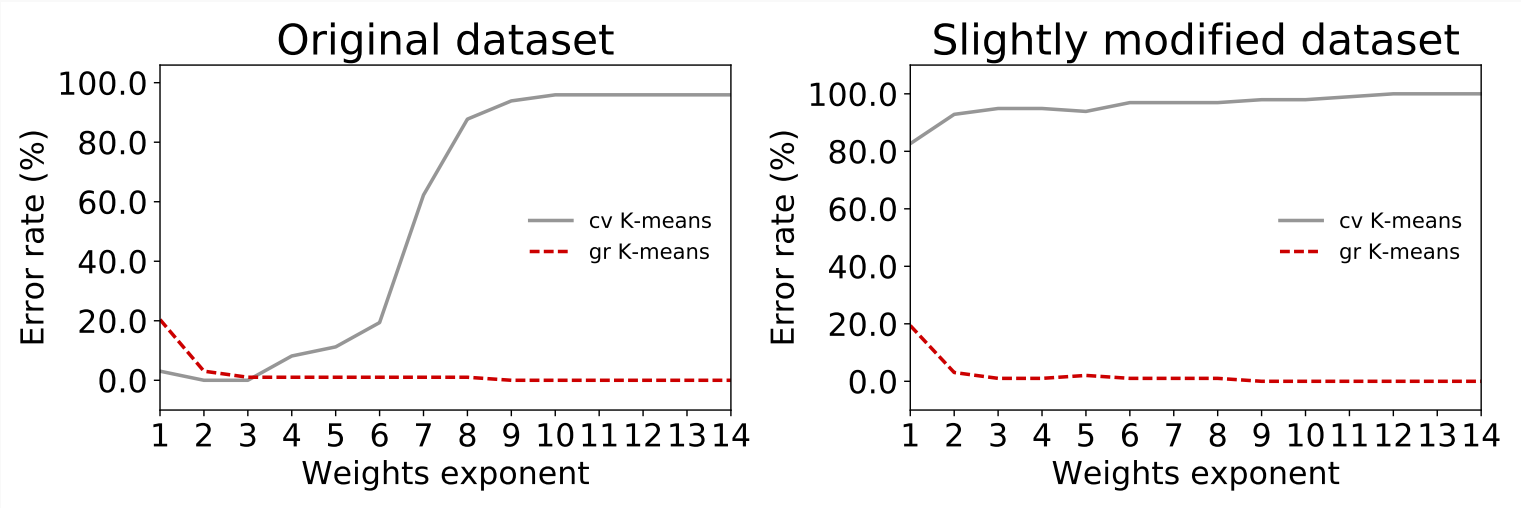
Results on 98 configurations



Robustness evaluation

Exponential weights

$$d(x_i, c_k) = \sqrt{\sum_{j=1}^{N} w_j^p (x_{ij} - c_{kj})^2}$$



Outline

Introduction

Preliminaries

Gap-Ratio K-means

Experimental Validation

Conclusion

Conclusion

- GR is a nice alternative to CV for data on interval different
- It appears to work well on the practical case studied here

Future Work

- Combine GR K means with data orthogonalization techniques
- Use more advance features for the application (pretrained CNN)