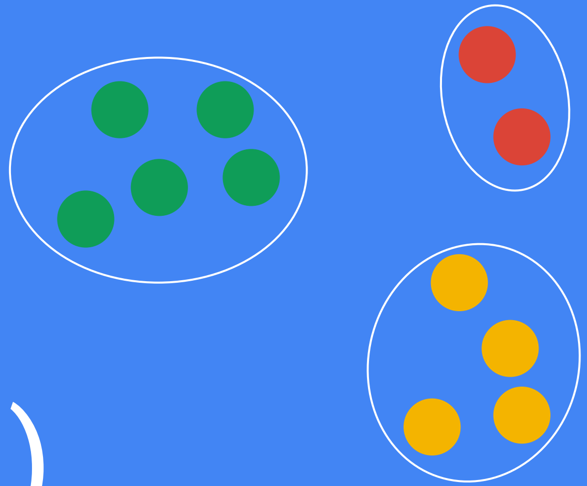
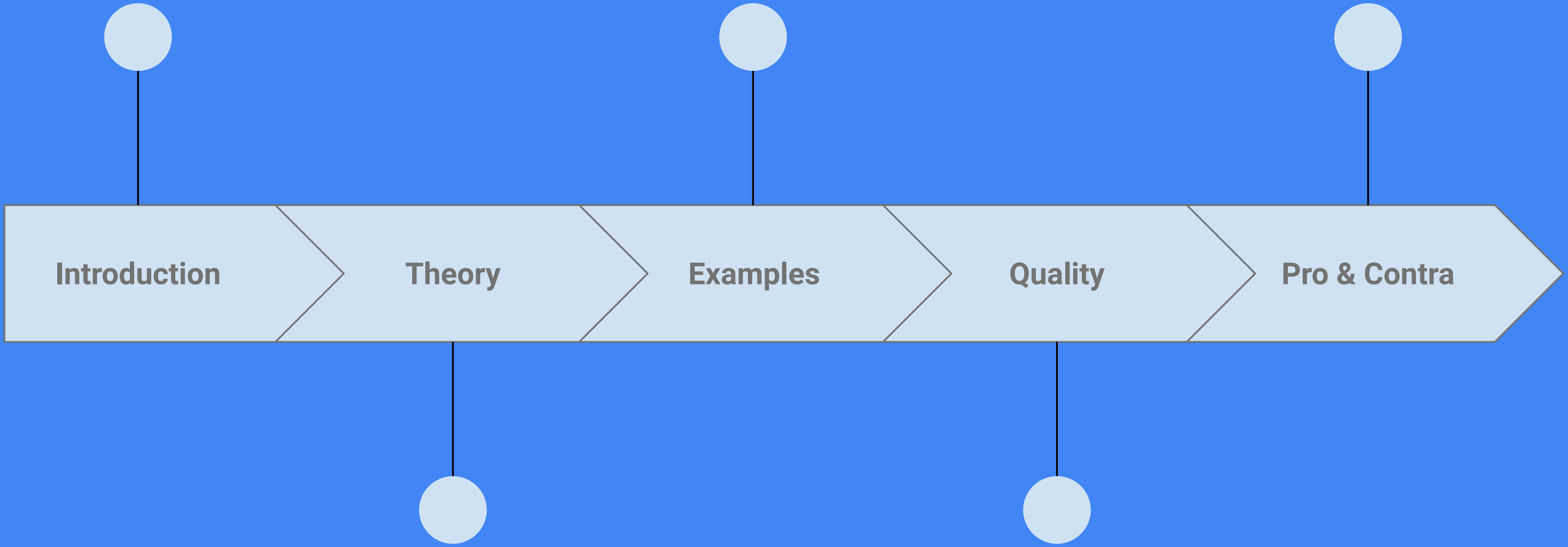
k-means (Clustering)



Knowledge Discovery (CENG-542)
Paper Presentation

Joel Amarou Heuer
IZTECH, 2022/04/26

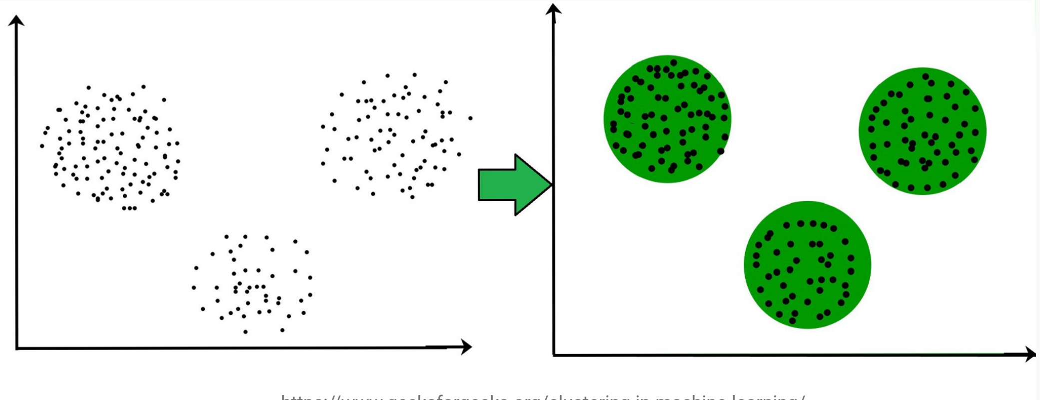
TOC (k-means)



Introduction

Motivation: to cluster given dataset

Question: How?



https://www.geeksforgeeks.org/clustering-in-machine-learning/

Answer: k-means algorithm (iterative way to **partition** dataset)

Approach: unsupervised machine learning (no training phase)

k-means (Theory)

Data: has *d* dimensions

Distance Measurement: euclidian

Nature of algorithm: greedy

Pseudocode (k-means)

(1) specify #clusters *k*

(2) randomly pick *k* data points as centroids

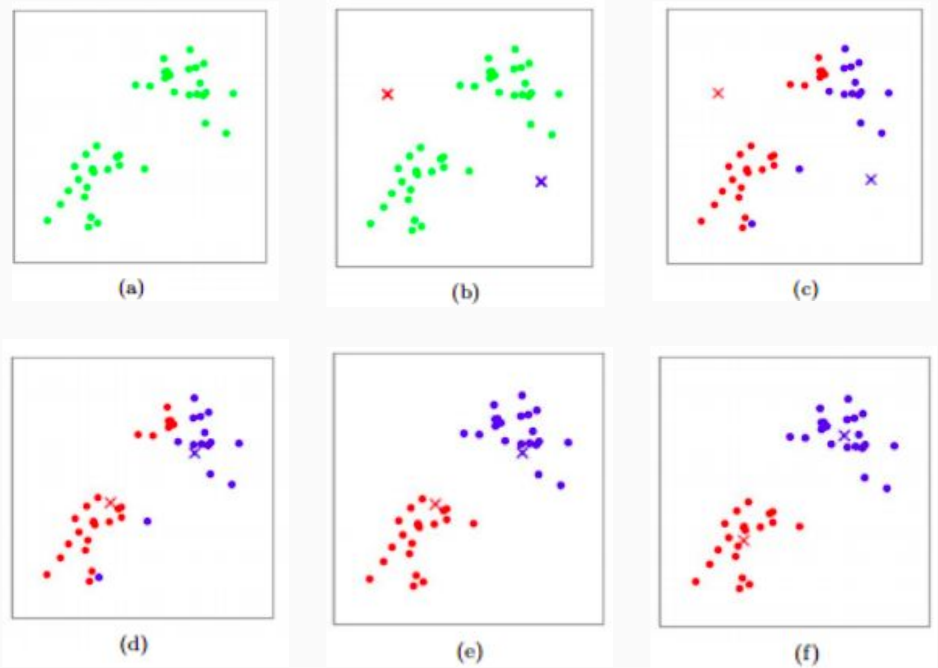
(3) **repeat**

(4) **[assignment]** **for each** data point *dp*: assign *dp* to closest centroid

(5) **[relocation]** **for each** cluster *c* : update *c*'s mean

(6) **until** centroids do not change

k-means – Example 1



https://stanford.edu/~cpiech/cs221/img/kmeansViz.png

Pseudocode (k-means)

(1) specify #clusters k

(2) randomly pick k data points as centroids
→ (b)

(3) **repeat**

(4) [assignment] **for each** data point dp: assign dp to closest centroid
→ (c), (e)

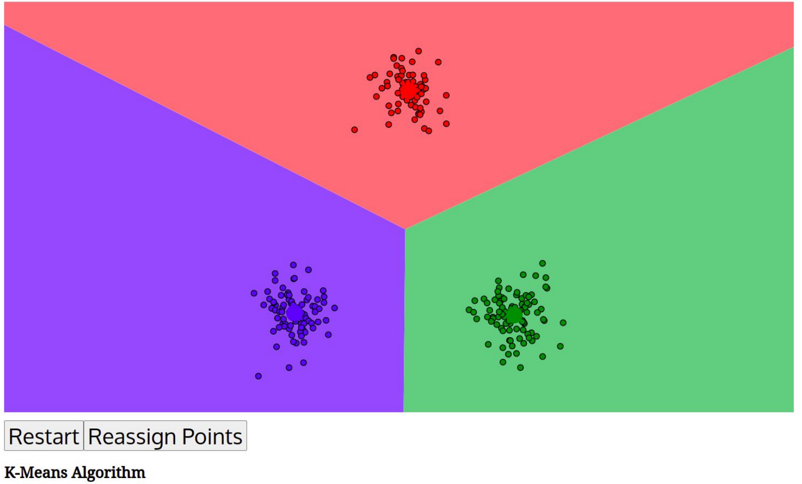
(5) [relocation] **for each** cluster c : update c's mean
→ (d), (f)

(6) **until** centroids do not change

k-means – Example 2 (interactive)

Experiment:

1. https://www.naftaliharris.com/blog/visualizing-k-means-clustering/
2. Press "*I'll Choose*"
3. Press "*Gaussian Mixture*"
4. Try different starting points & different *k*
   - k=6, 2 centroids for each cluster
   - k=6, 4 in A, 1 in B, 1 in C
   - k=3
   - k=2
   - k=1 (worst cluster)

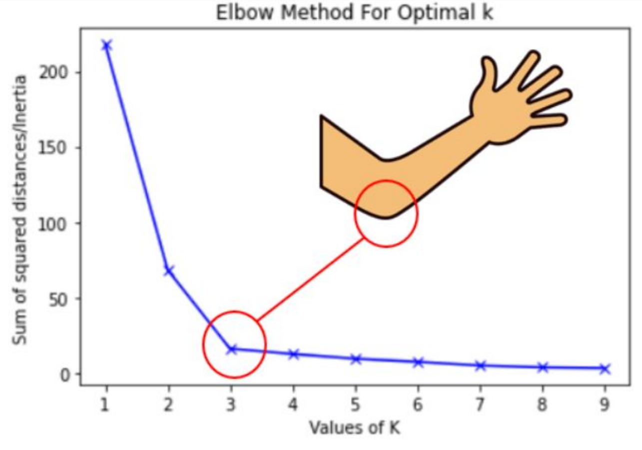


Quality

Quality measurement: within-cluster variation / within-cluster sum of squared error (WSS)

$$E = \sum_{\mathrm{i=1}}^{\mathrm{k}} \sum_{x \in \text{cluster } C_i} \text{dist}(x, \text{centroid of } C_i)$$

Elbow plot: plot WSS for different *k*



Line plot between K and inertia

https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

Pro & Contra

Pro	Contra
easy implementation	sensitive against outliers → *preprocess data* → *use k-medoids algorithm*
adapts to new data points	initial pick for centroids has impact on result → *repeat algorithm*
scalable (many data points)	affected from curse of dimensionality → *reduce dimensions (e. g. PCA)*
guaranteed convergence	not guaranteed to converge to the global optimum (often terminates at a local optimum) → *repeat algorithm*
runtime O(k * n * iterations)	manually choose *k* → *repeat algorithm & use elbow plot*

Thank you!

Questions?

About Me:

Joel Amarou Heuer
Freie Universität Berlin
Erasmus+ Student at IZTECH

joelheuer@iyte.edu.tr