Lecture 14:
K-means Clustering

Heidi Perry, PhD

Hack University

heidiperryphd@gmail.com

12/1/2016

Overview

Clustering

Cluster

A number of similar things that occur together (Merriam-Webster).

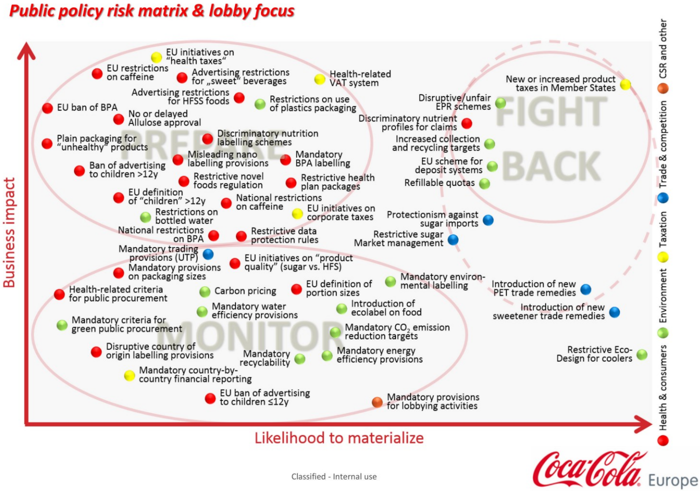
Clustering

Cluster

A number of similar things that occur together (Merriam-Webster).

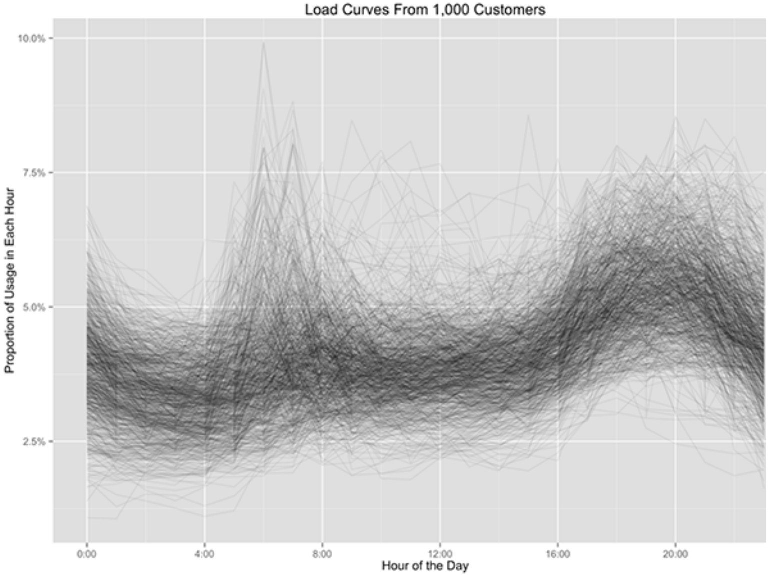
- Unsupervised learning - find patterns in unlabeled data
  - Uncover hidden structure in your dataset
  - Useful if you don't know what to look for
- Segment data into "similar" groups
  - Similarity measure is very important, but can be hard to define

Non-algorithmic Example



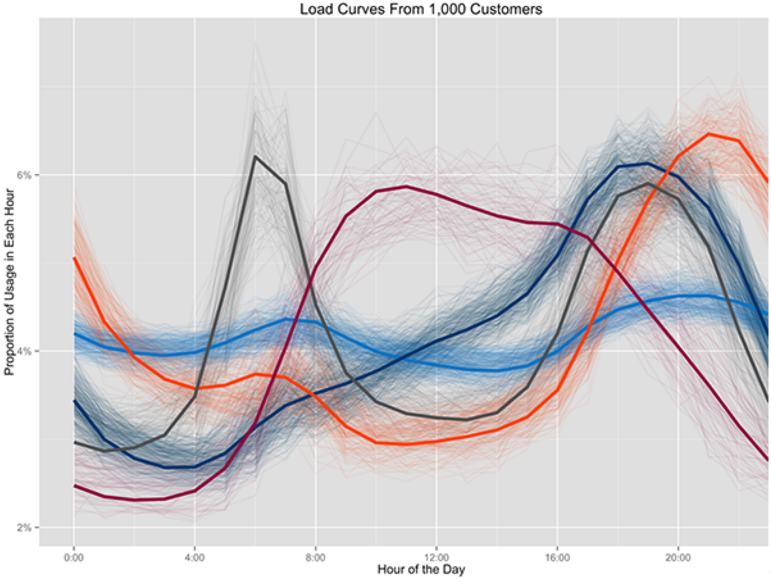
Downloaded from Ninja for Health, from leaked emails.

Time series clustering



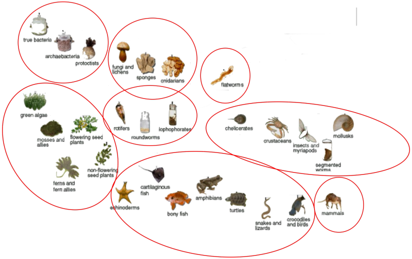
Source: Opower (2014)

Time series clustering



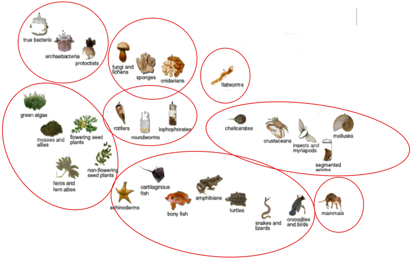
Clustering algorithms

Partition algorithms (flat)

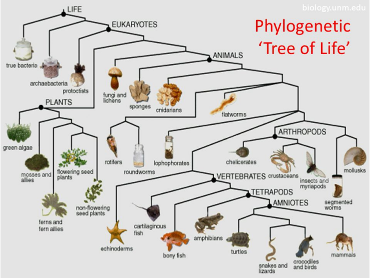


Clustering algorithms

Partition algorithms (flat)

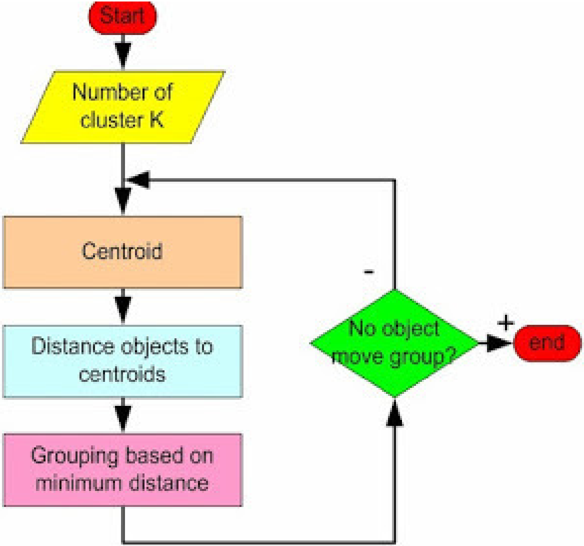


Hierarchical algorithms

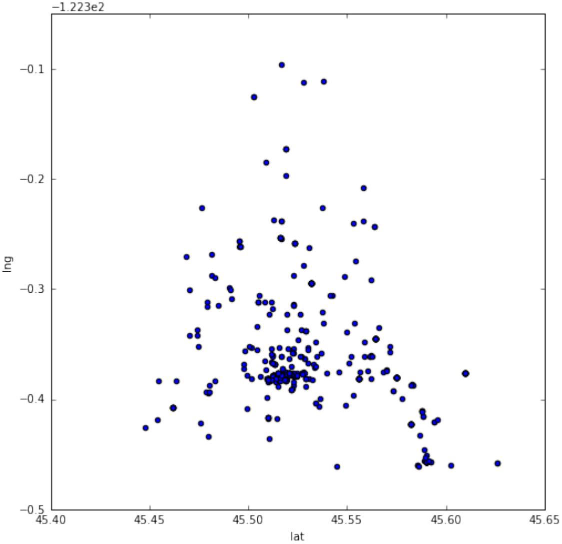


biology.unm.edu

K-means Algorithm

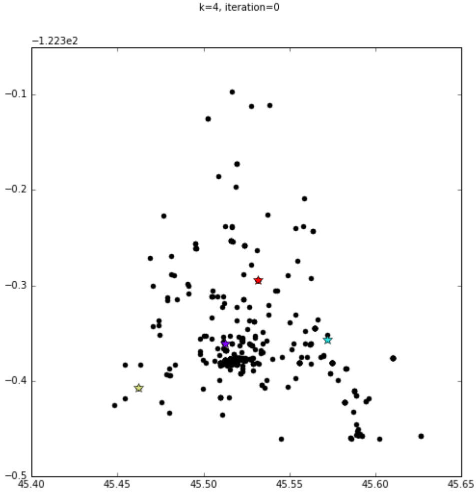


Example - Location of Public Art in Portland

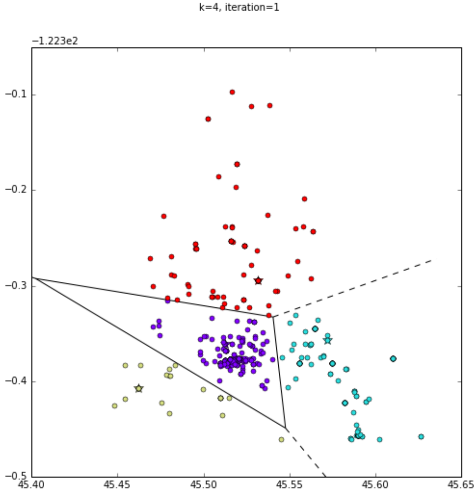


Location of Public Art in Portland

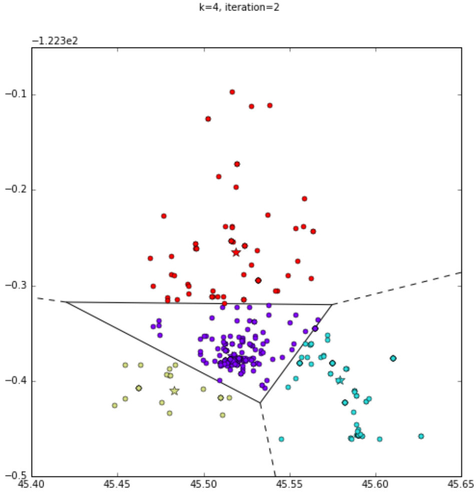
Example - Location of Public Art in Portland, iterations for k=4



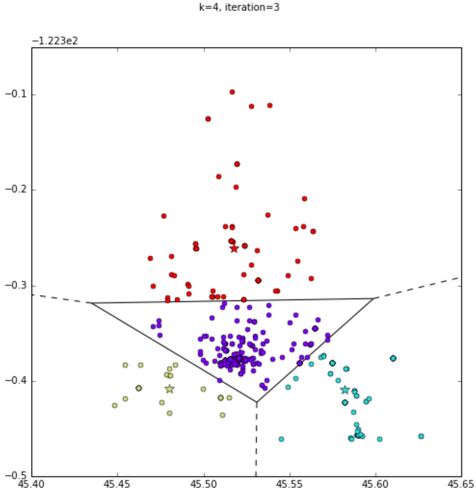
Example - Location of Public Art in Portland, iterations for k=4



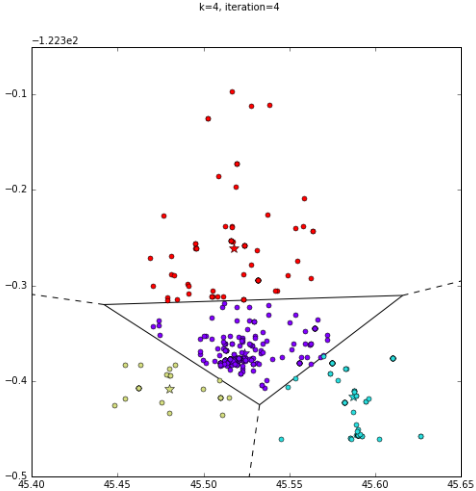
Example - Location of Public Art in Portland, iterations for k=4



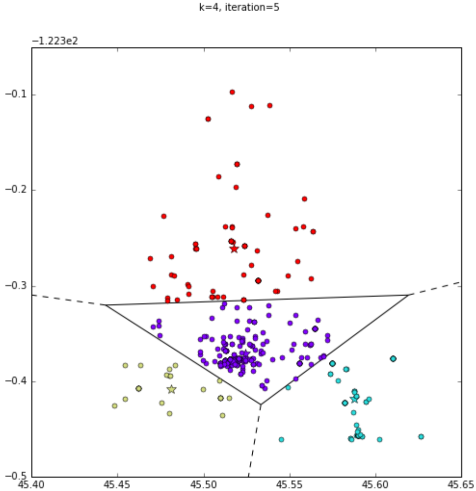
Example - Location of Public Art in Portland, iterations for k=4



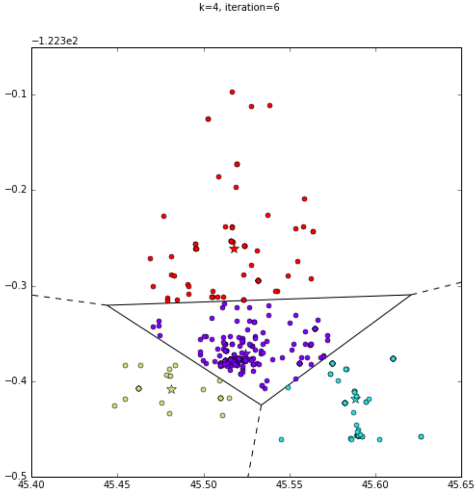
Example - Location of Public Art in Portland, iterations for k=4



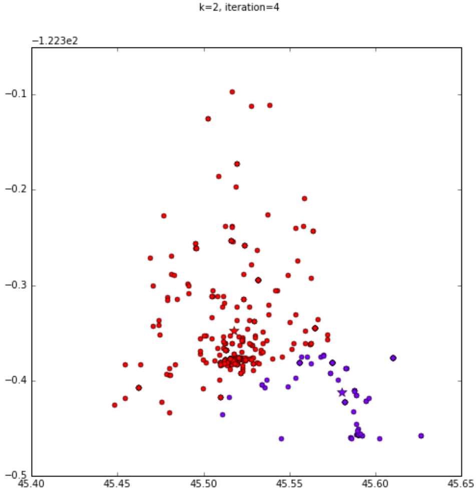
Example - Location of Public Art in Portland, iterations for k=4



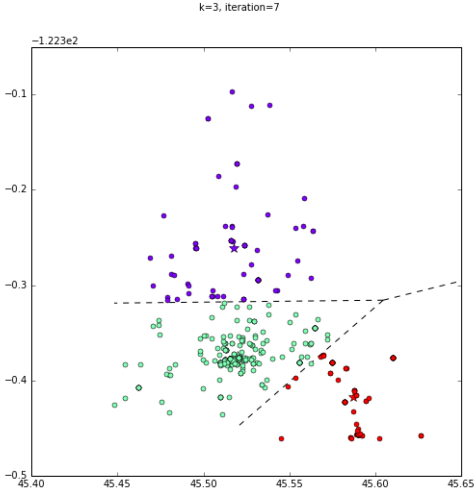
Example - Location of Public Art in Portland, iterations for k=4



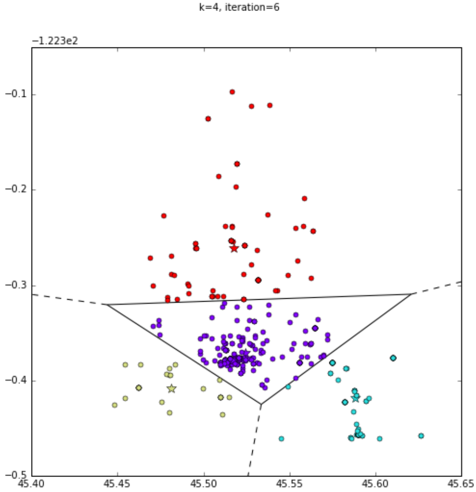
Example - Location of Public Art in Portland, various k



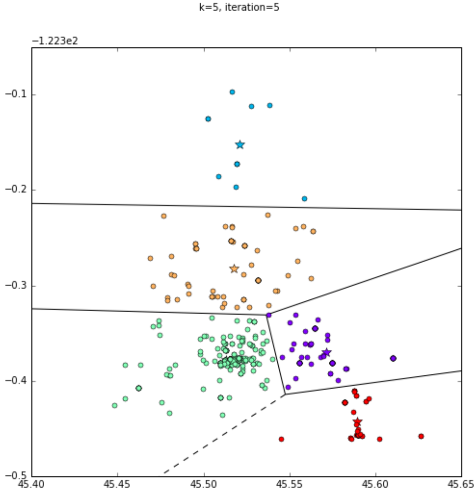
Example - Location of Public Art in Portland, various k



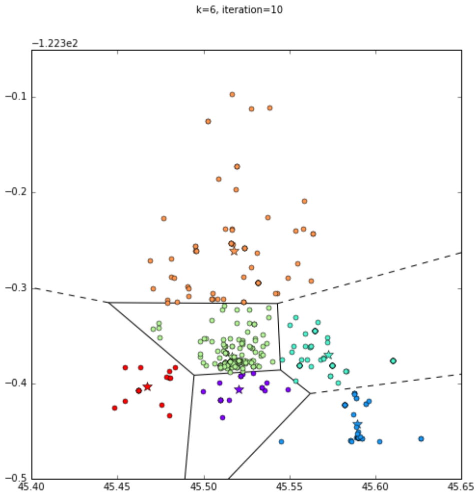
Example - Location of Public Art in Portland, various k



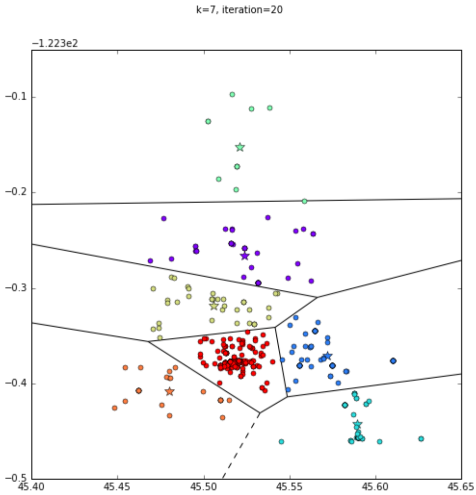
Example - Location of Public Art in Portland, various k



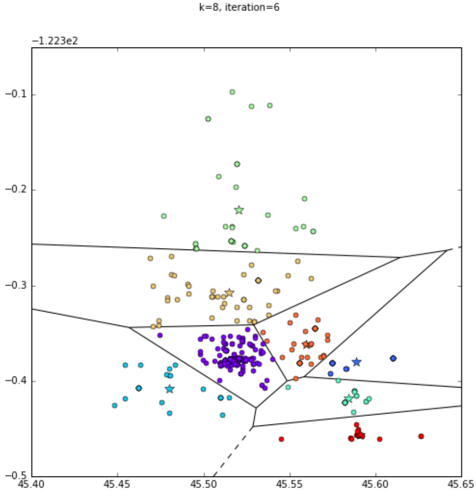
Example - Location of Public Art in Portland, various k



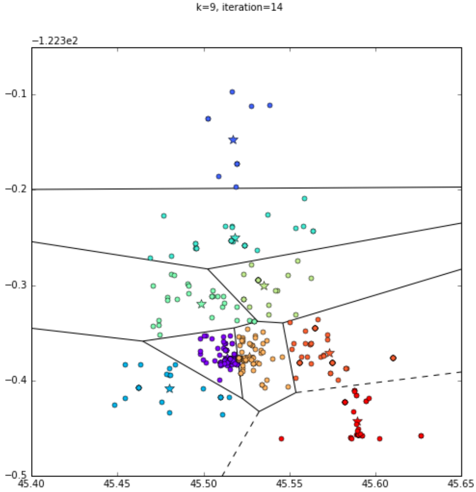
Example - Location of Public Art in Portland, various k



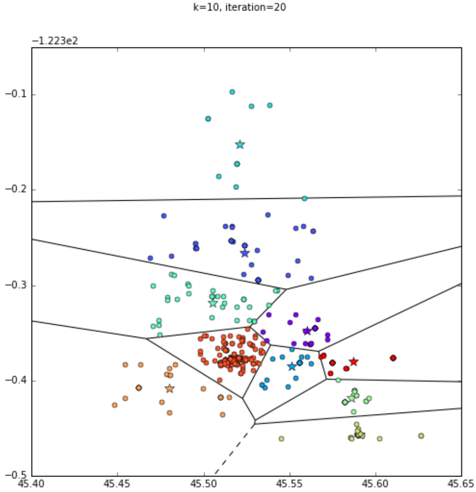
Example - Location of Public Art in Portland, various k



Example - Location of Public Art in Portland, various k



Example - Location of Public Art in Portland, various k



Objective Function

K-means algorithm minimizes the *residual sum of squares* of the data ($\vec{x_n}$) compared to the centroids ($\vec{\mu_k}$) of the cluster it belongs to ($r_{nk} = 1$ if $\vec{x_n}$ is in cluster k, else 0) for a given k:

$$J(\mu, r) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\vec{x_n} - \vec{(}\mu_k)||^2$$

Choosing k

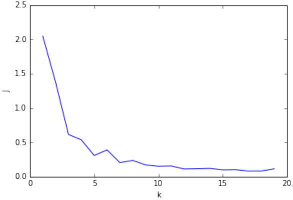
Objective Function

K-means algorithm minimizes the *residual sum of squares* of the data ($\vec{x_n}$) compared to the centroids ($\vec{\mu_k}$) of the cluster it belongs to ($r_{nk} = 1$ if $\vec{x_n}$ is in cluster k, else 0) for a given k:

$$J(\mu, r) = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} ||\vec{x_n} - \vec{(}\mu_k)||^2$$

One way to choose k

Find the "elbow point" in J vs. k (shown below for the public art example):



Defining 'Similar'

Similarity is inversely related to distance.

Defining ‘Similar’

Similarity is inversely related to distance.
Distance between what?

Defining 'Similar'

Similarity is inversely related to distance.
Distance between what?

- Text: Bag of words frequency vector

Defining 'Similar'

Similarity is inversely related to distance.
Distance between what?

- Text: Bag of words frequency vector
- Color: RGB levels (red, green, blue)

Defining 'Similar'

Similarity is inversely related to distance.

Distance between what?

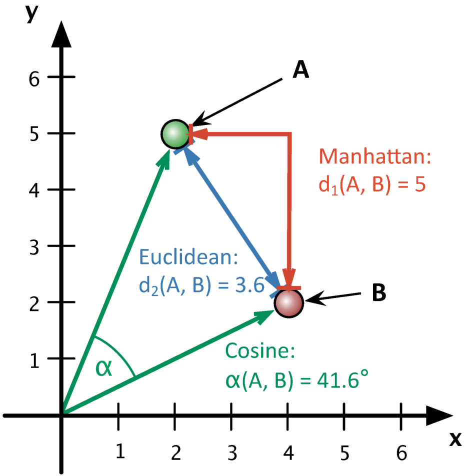
- Text: Bag of words frequency vector
- Color: RGB levels (red, green, blue)
- Generally: Any vector of numerical variables.

Defining 'Similar'

Similarity is inversely related to distance.
Distance between what?

- Text: Bag of words frequency vector
- Color: RGB levels (red, green, blue)
- Generally: Any vector of numerical variables.

Some distance measures:



[digitalhumanities]

K-means downfalls

- Sensitive to cluster center initialization.

K-means downfalls

- Sensitive to cluster center initialization.
- Only finds convex clusters

K-means downfalls

- Sensitive to cluster center initialization.
- Only finds convex clusters
- Does not handle clusters with different densities or sizes well

K-means downfalls

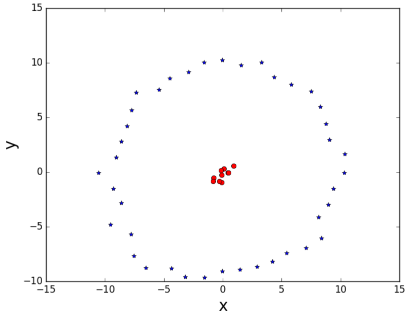
- Sensitive to cluster center initialization.
- Only finds convex clusters
- Does not handle clusters with different densities or sizes well
- Very sensitive to outliers

K-means downfalls

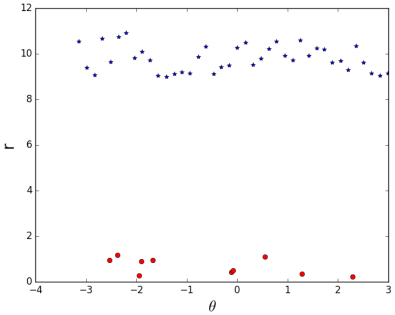
- Sensitive to cluster center initialization.
- Only finds convex clusters
- Does not handle clusters with different densities or sizes well
- Very sensitive to outliers
- See [Tan] pp 25-26 for examples.

Transforming can help

Some situations where k-means fails...



may be resolved with feature engineering.



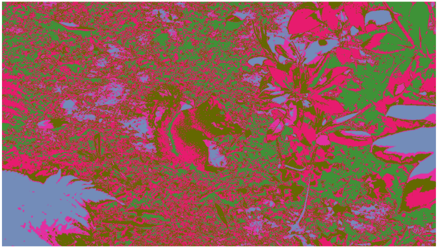
K-means to reduce colors in image

Original

K-means to reduce colors in image

Original

Cluster colors k=5



K-means to segment image

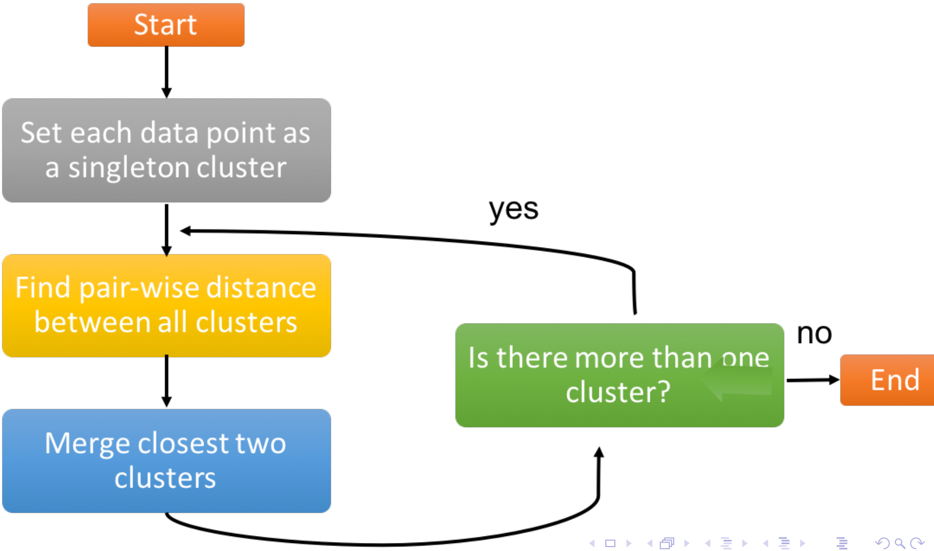
Original

K-means to segment image

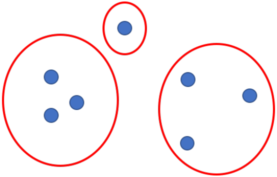
Original

Segment k=25

Agglomerative Clustering



Different choices result in different clustering behaviors.

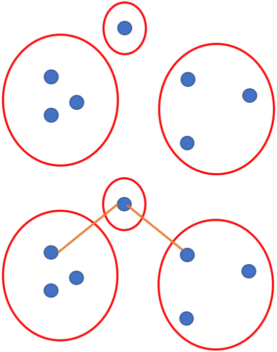


A few of the options:

- Closest pair (single-link clustering)
- Farthest pair (complete-link clustering)
- Average of all pairs

Defining Distance Between Clusters

Different choices result in different clustering behaviors.

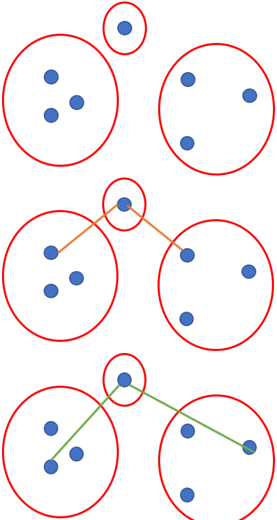


A few of the options:

- Closest pair (single-link clustering)
- Farthest pair (complete-link clustering)
- Average of all pairs

Defining Distance Between Clusters

Different choices result in different clustering behaviors.



A few of the options:

- Closest pair (single-link clustering)
- Farthest pair (complete-link clustering)
- Average of all pairs

References

- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
  Introduction to Data Mining

- Evert, S., Jannidis, F., Proisl, T., Vitt, T., Schch, C., Pielstrm, S., Reger, I. (2016). Outliers or Key Profiles? Understanding Distance Measures for Authorship Attribution. In Digital Humanities 2016: Conference Abstracts. Jagiellonian University & Pedagogical University, Krakw, pp. 188-191. Digital Humanities 2016

- Tibshirani, Robert; Walther, Guenther; and Hastie, Trevor (2001) Estimating the number of clusters in a data set via the gap statistic. J. R. Statist. Soc. B 63, Part 2, pp. 411-423. stanford.edu

Recommended Reading

Data Science from Scratch, Chapter 19

Further investigation

skikit learn Clustering

The Data Science Lab k-means clustering series