Lecture 14: K-means Clustering

Heidi Perry, PhD

Hack University heidiperryphd@gmail.com

12/1/2016

Overview

Introduction to Clustering

K-means algorithm

3 Hierarchical Clustering

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

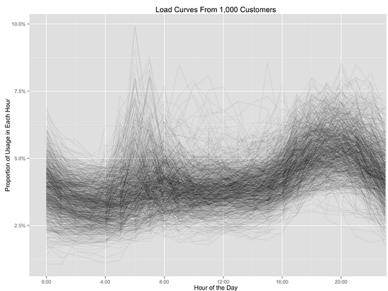
Non-algorithmic Example

Heidi Perry, PhD

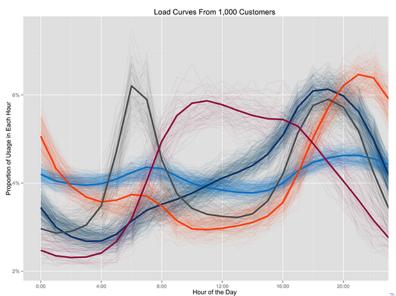
Public policy risk matrix & lobby focus EU initiatives on "health taxes" EU restrictions on caffeine Advertising restrictions for "sweet" beverages Health-related VAT system New or increased product taxes in Member States Advertising restrictions for HFSS foods Disruptive/unfair Restrictions on use EU ban of BPA of plastics packaging **EPR** schemes No or delayed Discriminatory nutrient Allulose approval profiles for claims Discriminatory nutrition Plain packaging for "unhealthy" products Increased collection labelling schemes and recycling targets Misleading nano labelling provisions BPA labelling EU scheme for Ban of advertising to children >12y deposit systems Restrictive novel Restrictive health **Business impact** Refillable quotas foods regulation plan packages EU definition of "children" >12y National restrictions on caffeine EU initiatives on Restrictions on Protectionism against corporate taxes bottled water sugar imports National restrictions on BPA Restrictive data Restrictive sugar Market management protection rules Mandatory trading provisions (UTP) EU initiatives on "product quality" (sugar vs. HFS) Mandatory provisions on packaging sizes Mandatory environ-EU definition of mental labelling Introduction of new Health-related criteria Carbon pricing portion sizes PET trade remedies for public procurement Introduction of Mandatory water Introduction of new ecolabel on food efficiency provisions sweetener trade remedies Mandatory criteria for Mandatory CO, emission green public procurement reduction targets Disruptive country of Mandatory Mandatory energy Restrictive Ecoorigin labelling provisions efficiency provisions Design for coolers recyclability 9 Mandatory country-bycountry financial reporting EU ban of advertising Mandatory provisions for lobbying activities Likelihood to materialize Classified - Internal use

Downloaded from Ninja for Health, from leaked emails.

Time series clustering



Time series clustering



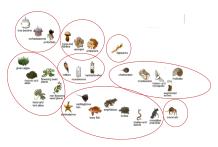
Clustering algorithms

Partition algorithms (flat)

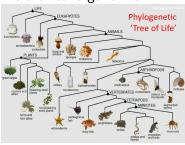


Clustering algorithms

Partition algorithms (flat)

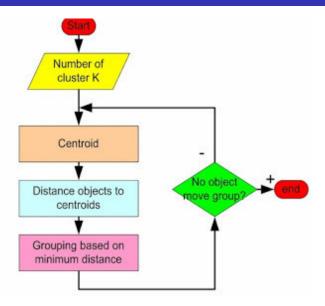


Hierarchical algorithms

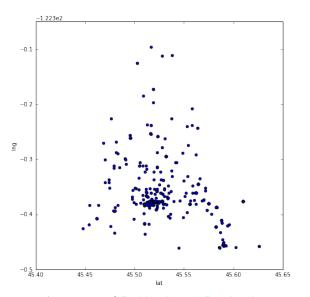


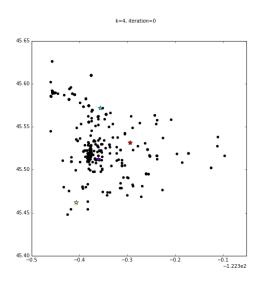
biology.unm.edu

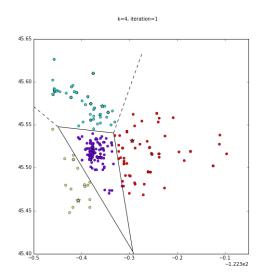
K-means Algorithm

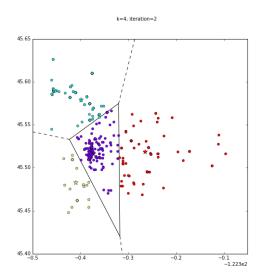


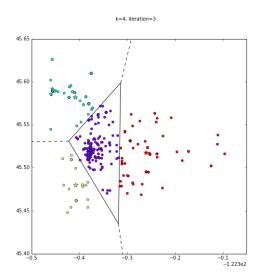
Example - Location of Public Art in Portland

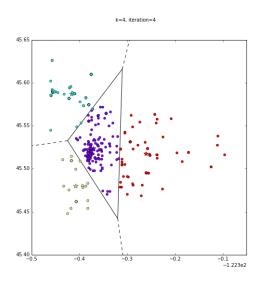


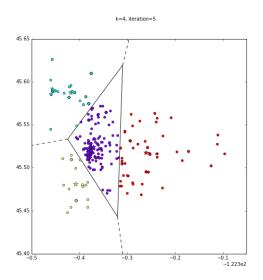


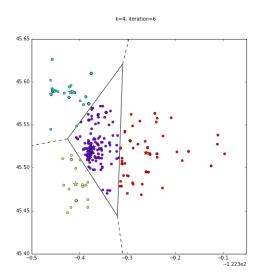


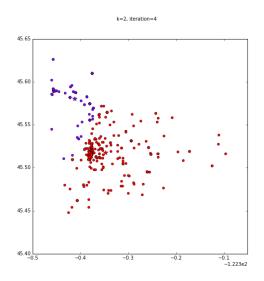


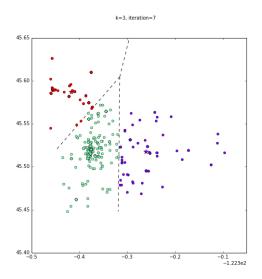


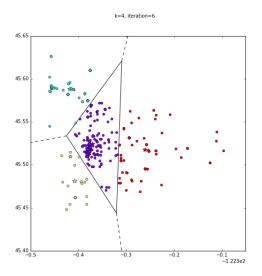


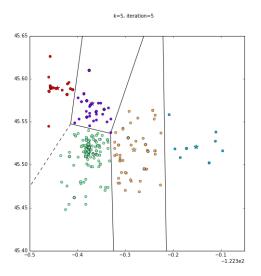


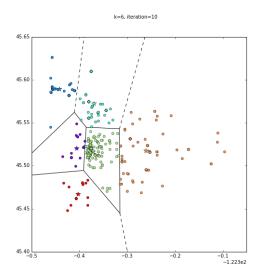


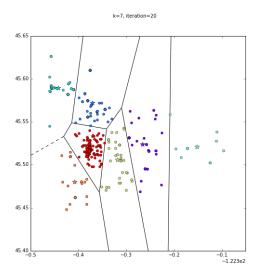


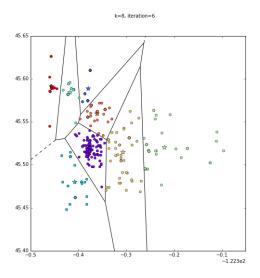


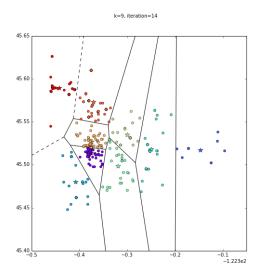


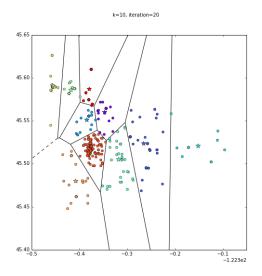












Choosing k

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 $(r_{nk}=1 \text{ if } \vec{x_n} \text{ is in cluster k, else 0})$ 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

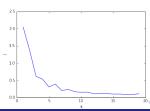
Objective Function

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 $(r_{nk}=1 \text{ if } \vec{x_n} \text{ is in cluster k, else 0})$ 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):



Heidi Perry, PhD

Similarity is inversely related to distance.

Similarity is inversely related to distance.

Distance between what?

Similarity is inversely related to distance.

Distance between what?

 Text: Bag of words frequency vector

Similarity is inversely related to distance.

Distance between what?

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

Similarity is inversely related to distance.

Distance between what?

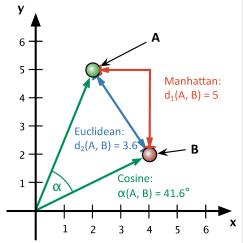
- Text: Bag of words frequency vector
- Color: RGB levels (red, green, blue)
- Generally: Any vector of numerical variables.

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.

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

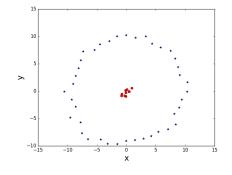
- Sensitive to cluster center initialization.
- Only finds convex clusters
- Does not handle clusters with different densities or sizes well

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

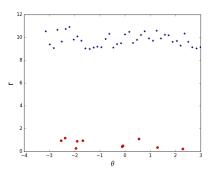
- Sensitive to cluster center initialization.
- Only finds convex clusters
- Does not handle clusters with different densities or sizes well
- Very sensitive to outliers
- See [Tan] pp 25-26 for examples.

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





Cluster colors k=10



K-means to segment image

Original



K-means to segment image

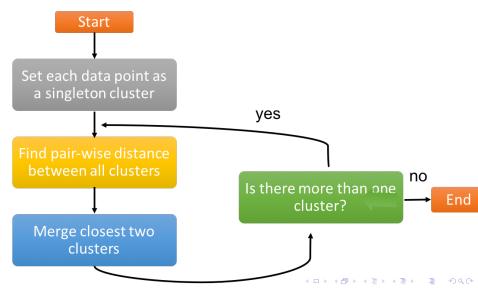
Original



Segment k=25

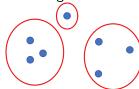


Agglomerative Clustering



Defining Distance Between Clusters

Different choices result in different clustering behaviors.



A few of the options:

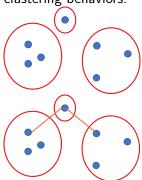
- Closest pair (single-link clustering)
- Farthest pair (complete-link clustering)
- Average of all pairs

Defining Distance Between Clusters

A few of the options:

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

Different choices result in different clustering behaviors.

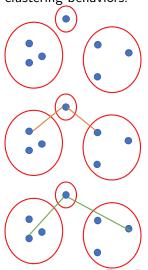


Defining Distance Between Clusters

A few of the options:

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

Different choices result in different clustering behaviors.



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

