


Learning from the data




Pattern recognition & data mining

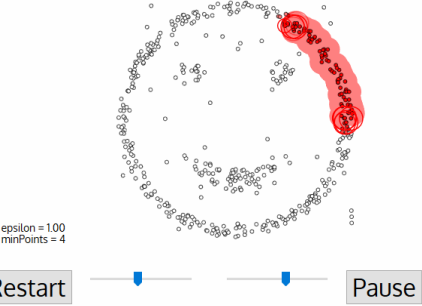
1

What is machine learning?
by Sabine Hauert, University of Bristol
(for the Royal Society)


<https://www.youtube.com/embed/F1wICerC40E>



2



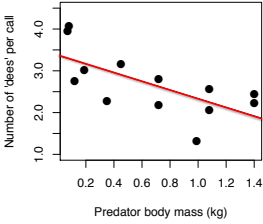
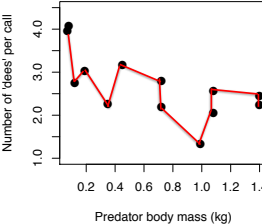
epsilon = 100
minPoints = 4

Restart  Pause

<https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68>

3

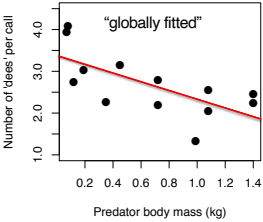
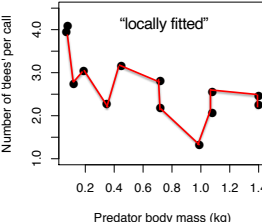
Learning from the data

What model do you prefer? Why?
Which model does better?

4

Learning from the data

What model do you prefer? Why?

“Intelligence is 10 million rules”
(Doug Lenat)...but Rules are meant to be
generalizable

5

Learning from the data - Machine learning algorithms

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed (Wikipedia).

6

**Learning from the data -
Machine learning algorithms**

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed (Wikipedia).
- **Machine learning** focuses on the development of computer algorithms that can change when exposed to new data. The process of **machine learning** is similar to that of data mining. The process is not strictly static following programming instructions; instead, they make data driven decisions (adapted from Wikipedia).

7

**Learning from the data -
Machine learning algorithms**

- **Machine learning** is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed (Wikipedia).
- **Machine learning** focuses on the development of computer algorithms that can change when exposed to new data. The process of **machine learning** is similar to that of data mining. The process is not strictly static following programming instructions; instead, they make data driven decisions (adapted from Wikipedia).
- Analysis based on **machine learning** may change when the learning process algorithm is run on the same data multiple times.

8

**Learning from the data -
Machine learning algorithms**

Machine learning mixes computer sciences and statistics and relaxes assumptions ("sometimes").

9

Learning from the data -
Machine learning algorithms

Foundations of
Machine Learning

Mehryar Mohri,
Afshin Rostamizadeh,
and Anand Tewari

10

Machine learning as an
inductive process

General

DEDUCTION vs **INDUCTION**

Theory
Hypothesis
Observation
Confirmation

Theory
Hypothesis
Pattern
Observation

Specific

ARISTOTLE

SHERLOCK

Specific

<https://danielmiessler.com/blog/the-difference-between-deductive-and-inductive-reasoning/>

11

Learning from the data -
Machine learning algorithms

Model level

model

Induction phase (specific to general; i.e., looking for a pattern in data and then generalize it)

deduction phase (general to specific)

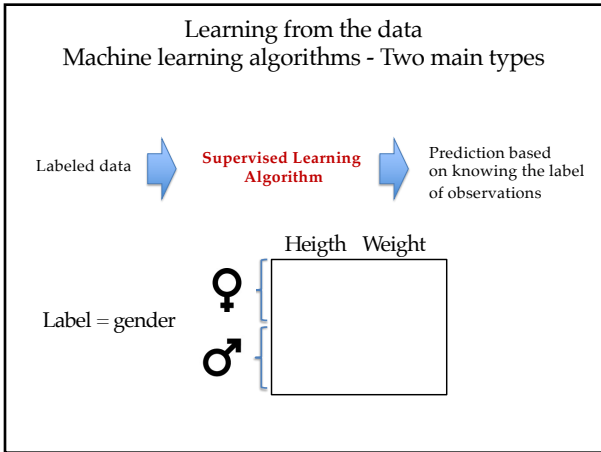
Data level

training data

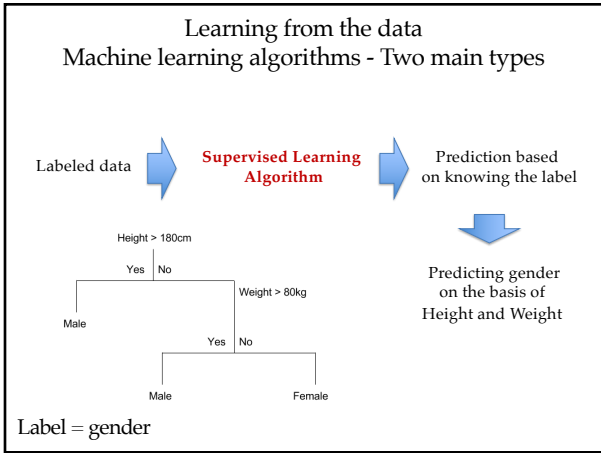
prediction & validation

Modified from
<http://www.cs.joensuu.fi/~whamala/skz/ml.html>

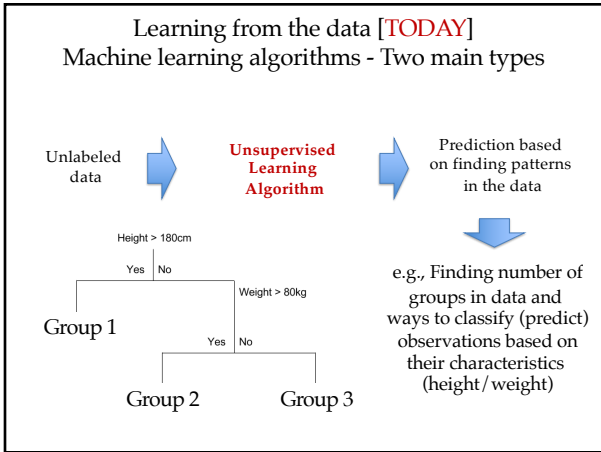
12



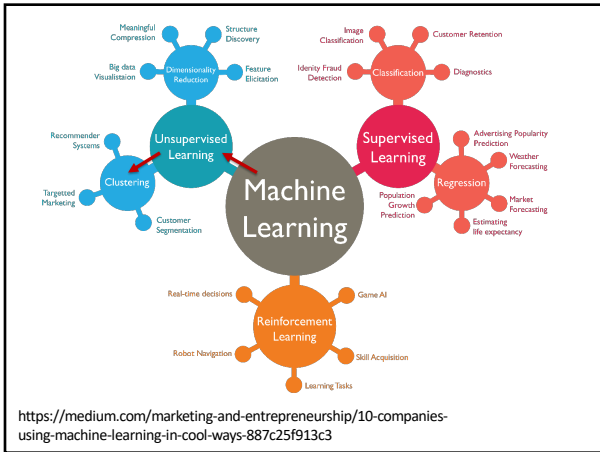
13



14



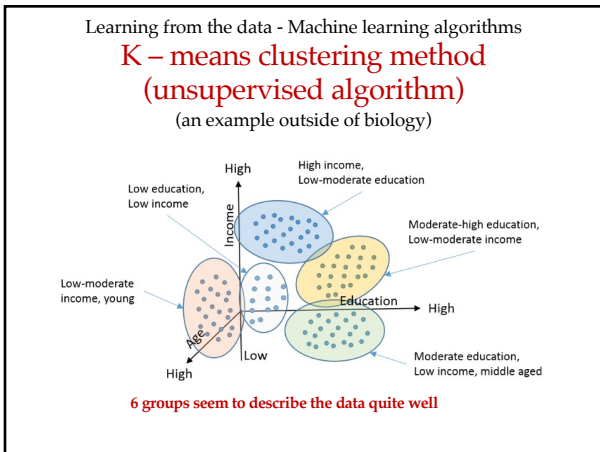
15



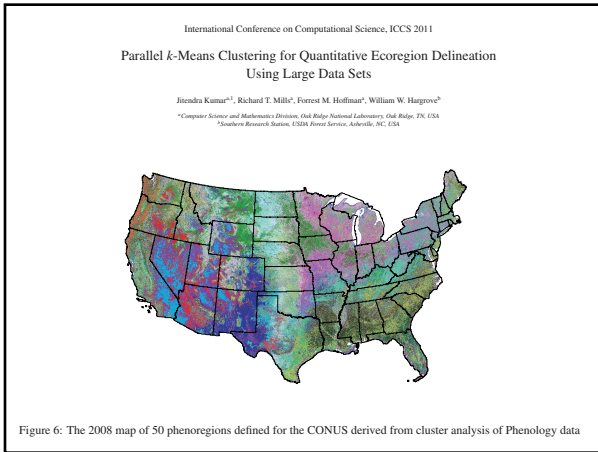
16

The k-means clustering algorithm
Easy to see what it does (video)
<https://www.youtube.com/watch?v=4b5d3muPQmA>

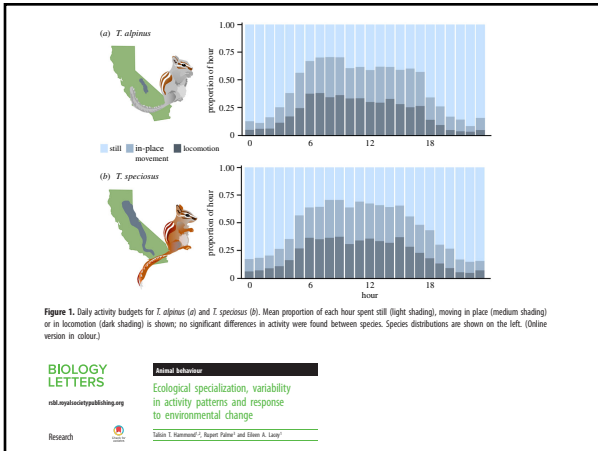
17



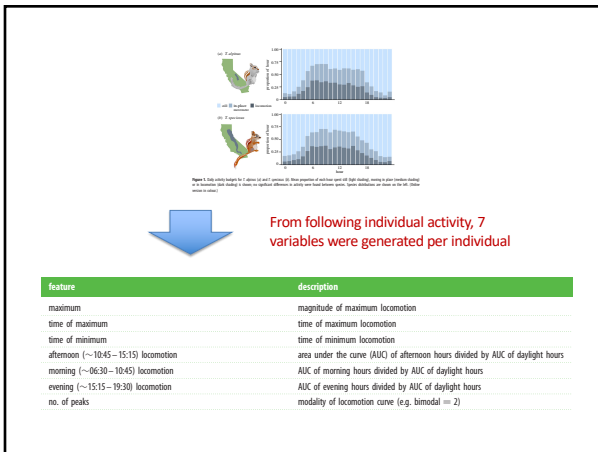
18



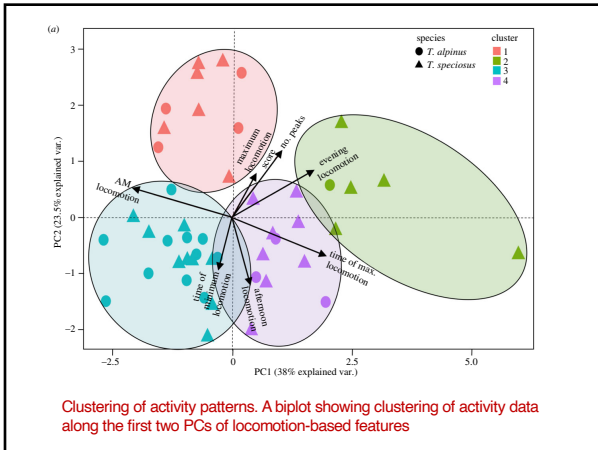
19



20



21



22

The basis of k-means

- Partition n points (observations) across multiple variables into k groups.
- The goal is to minimize an objective function (here the sum-of-squares of multivariate distances (Euclidean) within groups).

$$\text{objective function} \leftarrow J = \sum_{j=1}^k \sum_{t=1}^n \|x_t^{(j)} - c_j\|^2$$

number of clusters number of cases centroid for cluster j
Distance function

23

Learning from the data – Machine learning algorithms: k – means

We will consider only two dimensions here for visual simplicity (Height and Weight)

Unlabeled data

➔

Unsupervised Learning Algorithm

➔

Prediction based on finding patterns in the data

Height > 180cm

Yes | No

Group 1

Weight > 80kg

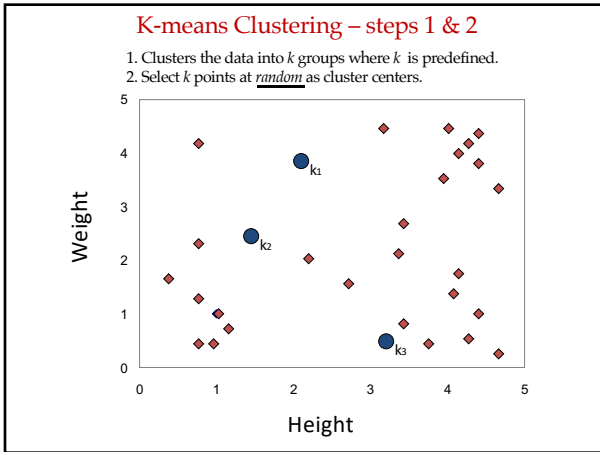
Yes | No

Group 2 | Group 3

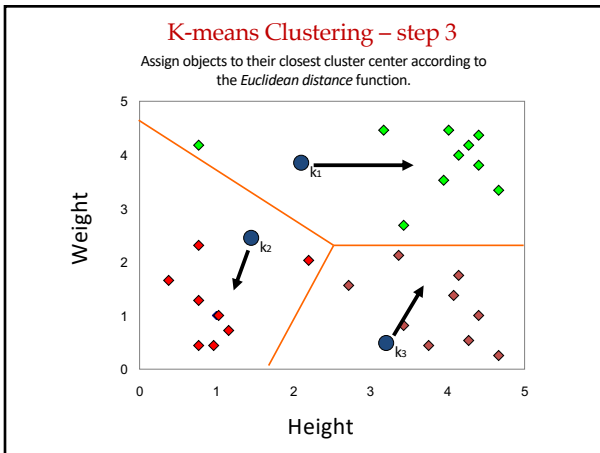
➔

e.g., Finding number of groups in data and ways to classify (predict) observations based on their characteristics (height/weight)

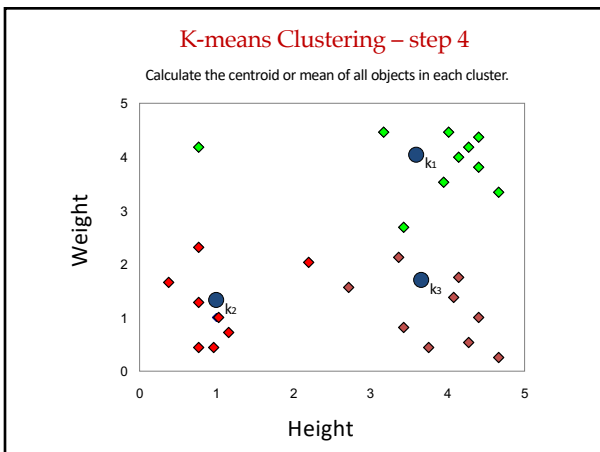
24



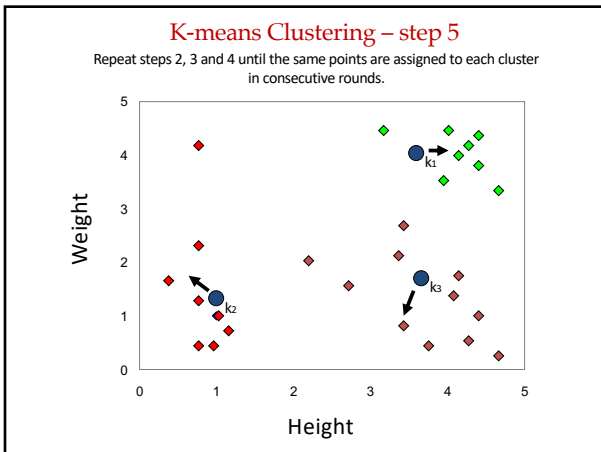
25



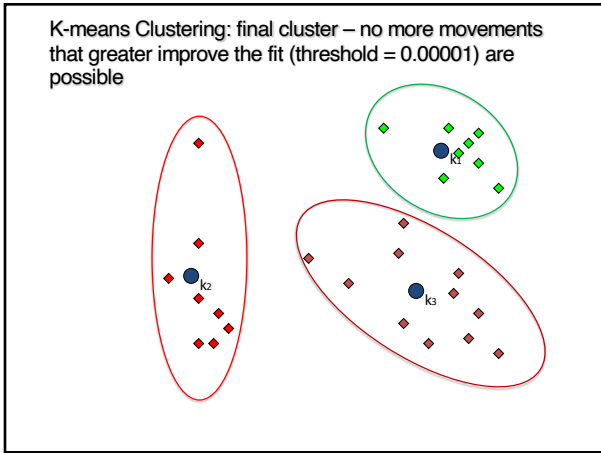
26



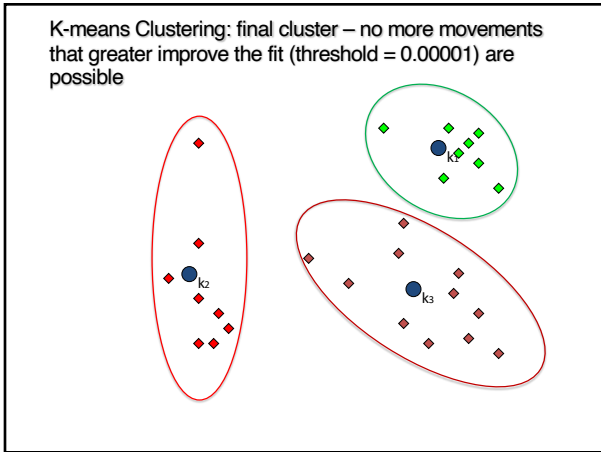
27



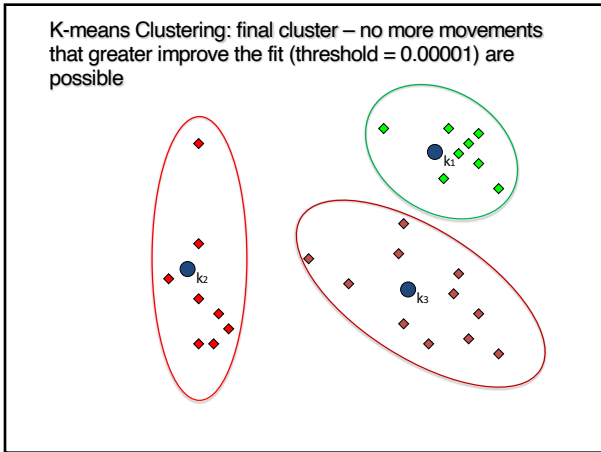
28



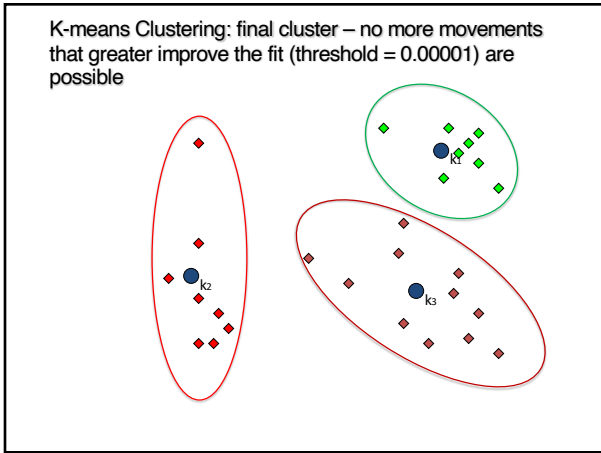
29



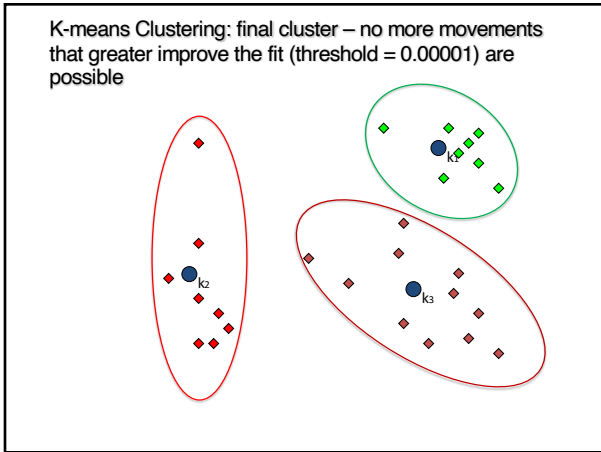
30



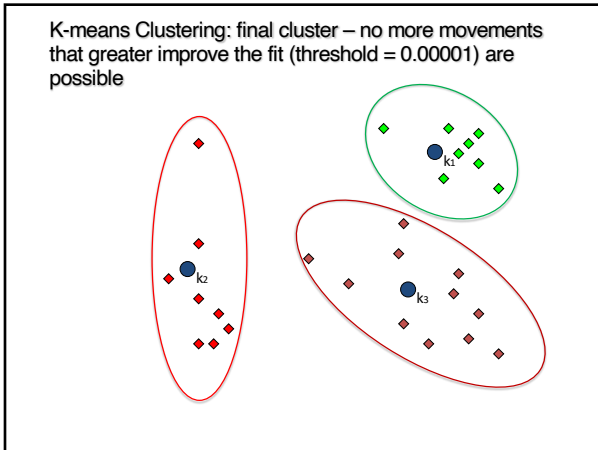
31



32



33



34

The (iterative) k-means algorithm (summary of “general” algorithm – there are others)

The number of clusters, k , is decided first; the iterative steps are then:

- 1) Generate an initial set of k points as the first estimate of the cluster points (random seed points).
- 2) Loop over all observations reassigning them to the group with the closest mean value.
- 3) Re-compute the mean of each group.

Iterate steps 2 and 3 until convergence (i.e., the mean distance of each object to its group mean does not change according to a very small threshold (e.g., 0.000001)).

35

The (iterative) k-means algorithm (summary of “general” algorithm – there are others)

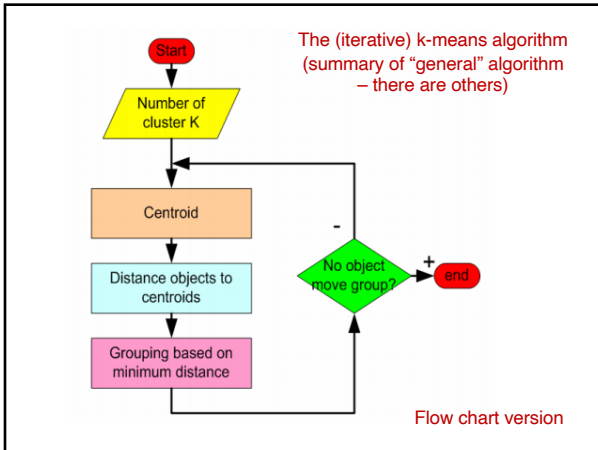
The number of clusters, k , is decided first; the iterative steps are then:

- 1) Generate an initial set of k points as the first estimate of the cluster points (random seed points).
- 2) Loop over all observations reassigning them to the group with the closest mean value. Assign objects to their closest cluster center according to the *Euclidean distance* function.
- 3) Re-compute the mean (multivariate centroids) of each group.

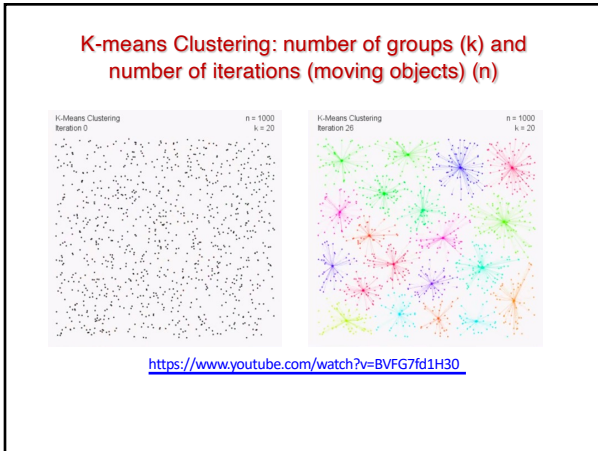
Iterate steps 2 and 3 until convergence (i.e., the mean distance of each object to its group mean does not change according to a very small threshold (e.g., 0.000001)).

An **iterative method** is called convergent if the corresponding sequence converges regardless of the initial approximations (random seed points).

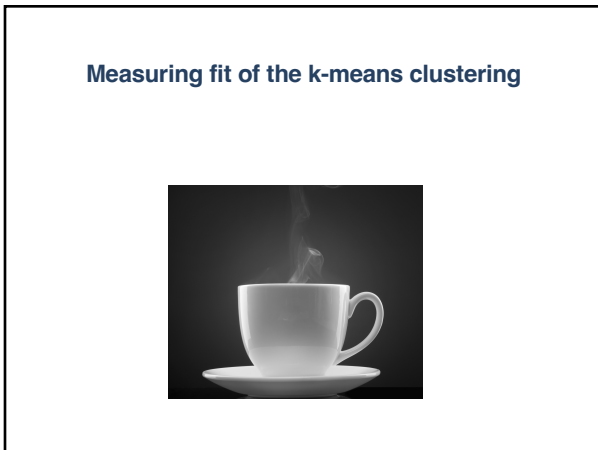
36



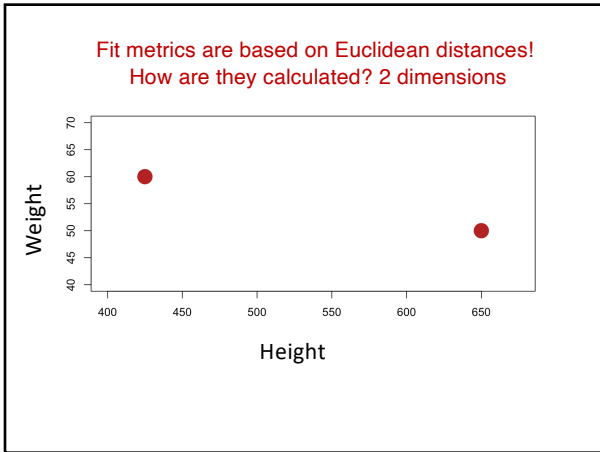
37



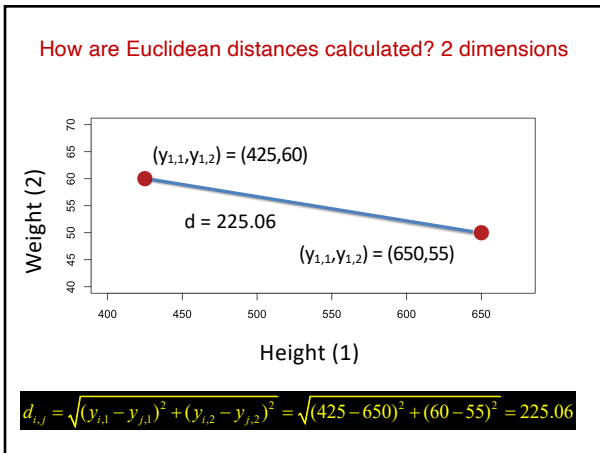
38



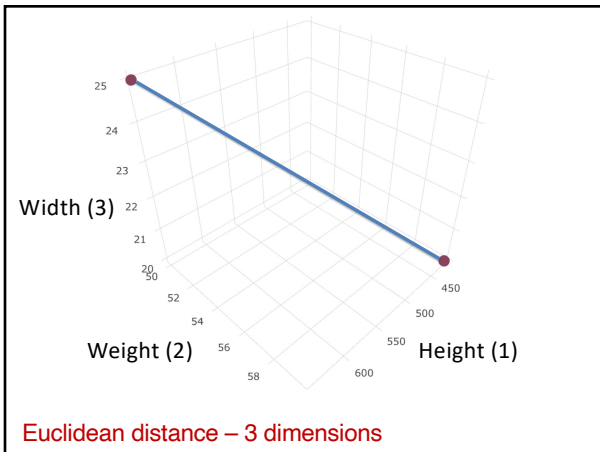
39



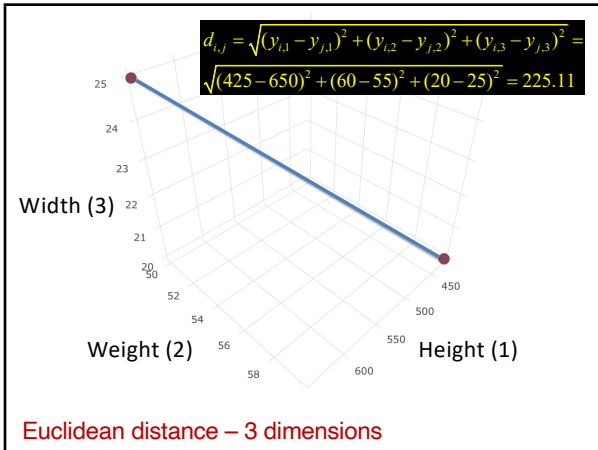
40



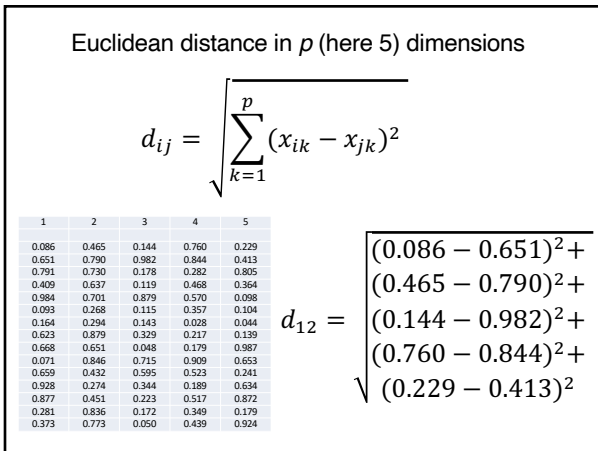
41



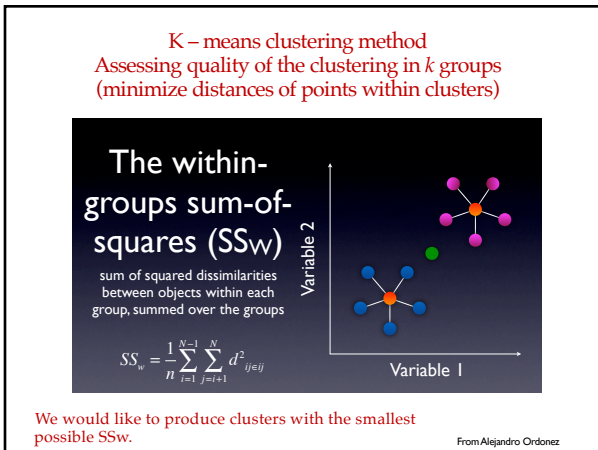
42



43



44



45

What is the optimal number of group?
lots of methods, e.g., the “elbow method”

SSw = Average within cluster distance to centroid

The within-groups sum-of-squares (SS_w)
sum of squared dissimilarities between objects within each group, summed over the groups

$$SS_w = \frac{1}{n} \sum_{j=1}^{n-1} \sum_{i=j+1}^n d_{ij}^2$$

Variable I
Variable 2

46

K – means clustering method
Quality of the clustering in k groups : SS_A/SS_T

The between-groups sum-of-squares (SS_A)
sum of squared dissimilarities between group means and the overall mean. It can be determined from the usual additive partitioning of the SS as described for ANOVA

$$SS_A = SS_T - SS_w$$

The total sum-of-squares (SS_T)
the sum of squared dissimilarities between all pairs of objects divided by N

$$SS_T = \frac{1}{N} \sum_{i=1}^{N-1} \sum_{j=i+1}^N d_{ij}^2$$

The SS_A/SS_T % is a measure of the total variance in the data set that is explained by the clustering. k-means minimize the within group dispersion and maximize the between-group dispersion. By assigning the samples to k clusters rather than n (number of samples) clusters achieved a reduction in sums of squares of SS_A/SS_T %.

47

What is the optimal number of group?
lots of methods, e.g., total variance explained

total variance explained (SS_A/SS_T)

Number of groups K

48

K – means clustering method

Quality of the clustering in k groups : SSI

The between-groups sum-of-squares (SS_B)

sum of squared dissimilarities between group means and the overall mean. It can be determined from the usual additive partitioning of the SS, as described for ANOVA.

$$SS_B = SS_T - SS_W$$

The within-groups sum-of-squares (SS_W)

sum of squared dissimilarities between objects within each group, summed over the groups.

$$SS_W = \frac{1}{N} \sum_{i=1}^k \sum_{j=1}^{n_i} d_{ij}^2$$

simple structure index (SSI) =

$$(SS_B / (K-1)) / (SS_W / (n-K))$$

n = number of objects (observations, data points); k = number of groups

49

K-means as a predictive model: for each of new observations we can estimate its probability to belonging to a particular group (cluster)

50

Linear versus non-linear group partitioning

51

K-means is used in a variety of problems

IJCAT International Journal of Computing and Technology, Volume 1, Issue 4, May 2014
ISSN : 2348 - 6090
www.IJCAT.org

Human Genome Data Clustering Using K-Means Algorithm

¹Amrita A. Kulkarni, ²Prof. Deepak Kaggate

¹Department of C.S.E., GHRAET, Nagpur University, Nagpur, Maharashtra, India


²Department of C.S.E., GHRAET, Nagpur University, Nagpur, Maharashtra, India

52

K-means is used in a variety of problems


Biomolecular Detection and Quantification 13 (2017) 7–31

Contents lists available at ScienceDirect



Biomolecular Detection and Quantification


journal homepage: www.elsevier.com/locate/bdq




Research paper

***K-means and cluster models for cancer signatures**

Zura Kakushadze^{a,b,1,*}, Willie Yu^c






COMPUTATIONAL GENOMICS APPROACHES TO PRECISION MEDICINE


53

K-means is used in a variety of problems


Methods in Ecology and Evolution




Volume 4, Issue 6
June 2013
Pages 542–551

Research Article  Free Access

Spherical k-means clustering is good for interpreting multivariate species occurrence data

Mark O. Hill , Colin A. Harrower, Christopher D. Preston

First published: 2 April 2013 | <https://doi.org/10.1111/2041-210X.12038> | Cited by:3




Frontiers of Environmental Science & Engineering

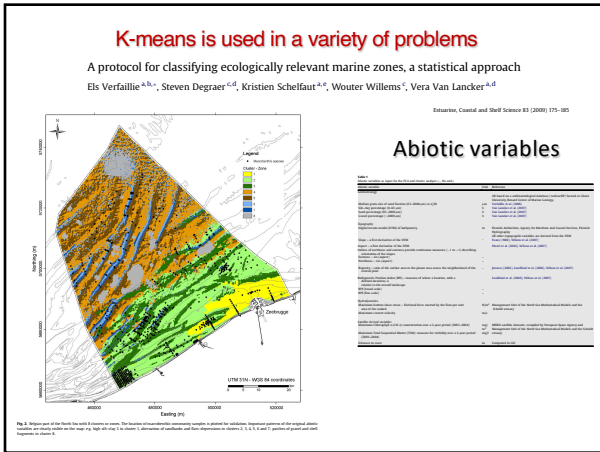
February 2014, Volume 8, Issue 1, pp 117–127 | [Cite as](#)

Application of *k*-means clustering to environmental risk zoning of the chemical industrial area

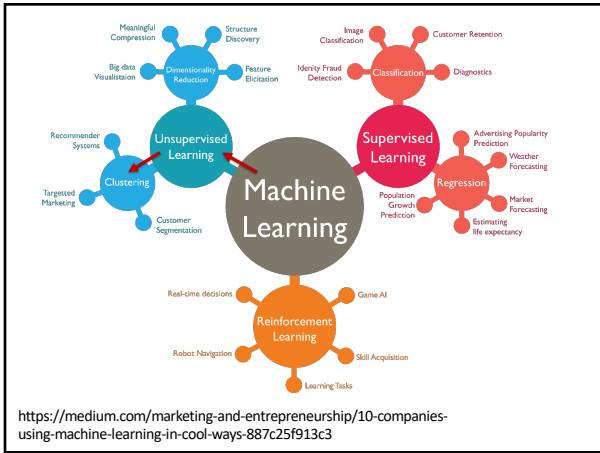
Authors [Authors and affiliations](#)

Weifang Shi, Weihua Zeng 

54



55



56
