FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

NATURWISSENSCHAFTLICHE
FAKULTÄT

Introduction

Lecture “Mathematics of Learning (Maths of Data Science) 1”

Frauke Liers
Friedrich-Alexander-Universität Erlangen-Nürnberg
Winter Semester 2021 / 22
Oct 20th, 2021

DISCRETE
OPTIMIZATION

Preliminaries

- *Many thanks to Martin Burger, Daniel Tenbrinck, and Philipp Wacker for allowing me to use the material they have compiled in earlier lectures!!*
- We will start from this for the first chapter, then adapt/edit, etc., according to needs.

Preliminaries

- *Many thanks to Martin Burger, Daniel Tenbrinck, and Philipp Wacker for allowing me to use the material they have compiled in earlier lectures!!*
- We will start from this for the first chapter, then adapt/edit, etc., according to needs.
- lecture: each Wednesday 8:15 am - 9:45 German time, H13, hybrid and video streaming. Recordings will be available also later in studon course.
- Chat in studon MoL course can be used during lecture, will be read regularly.

Preliminaries

- *Many thanks to Martin Burger, Daniel Tenbrinck, and Philipp Wacker for allowing me to use the material they have compiled in earlier lectures!!*
- We will start from this for the first chapter, then adapt/edit, etc., according to needs.
- lecture: each Wednesday 8:15 am - 9:45 German time, H13, hybrid and video streaming. Recordings will be available also later in studon course.
- Chat in studon MoL course can be used during lecture, will be read regularly.
- Exercises directly afterwards in person, and/or online at 11 AM, link see studon. (Florian Rösel)

Preliminaries

- Exercises consist of both mathematical and implementation tasks

Preliminaries

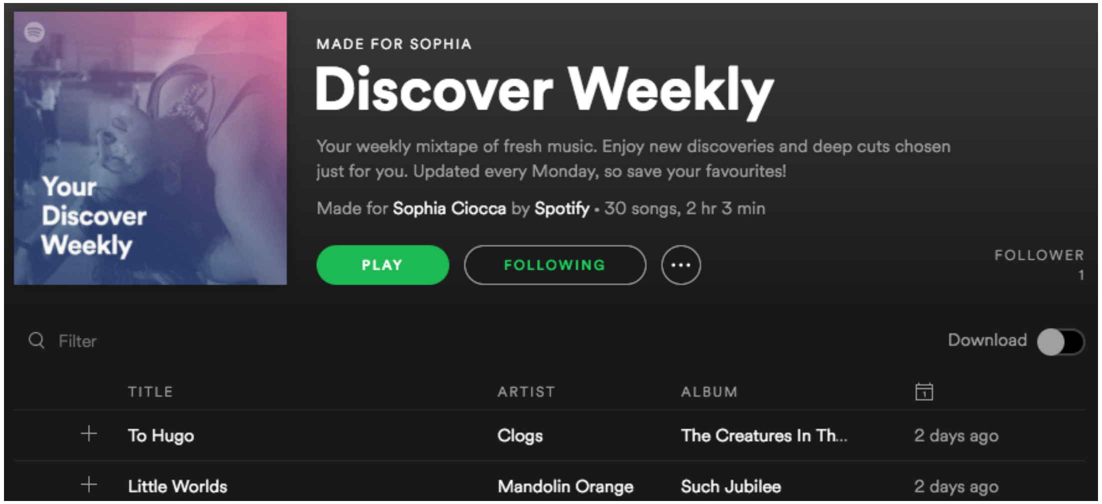
- Exercises consist of both mathematical and implementation tasks
- Please fill out the anonymous survey on background knowledge from studon course. (see studon section Tell us about your knowledge...)
- Students need to pass a written exam at the end of the semester.
- Passing this course is mandatory for Master Data Science.

What is Data Science?

- relatively **new field** due to recent boost in digital transforation $\rightarrow$ *digitization, IoT, paperless office, ...*
- needs mathematics, computer science, and domain knowledge
- well-known examples for successfully harvesting and interpreting data:
  $\rightarrow$ Google
  $\rightarrow$ Spotify
  $\rightarrow$ Amazon

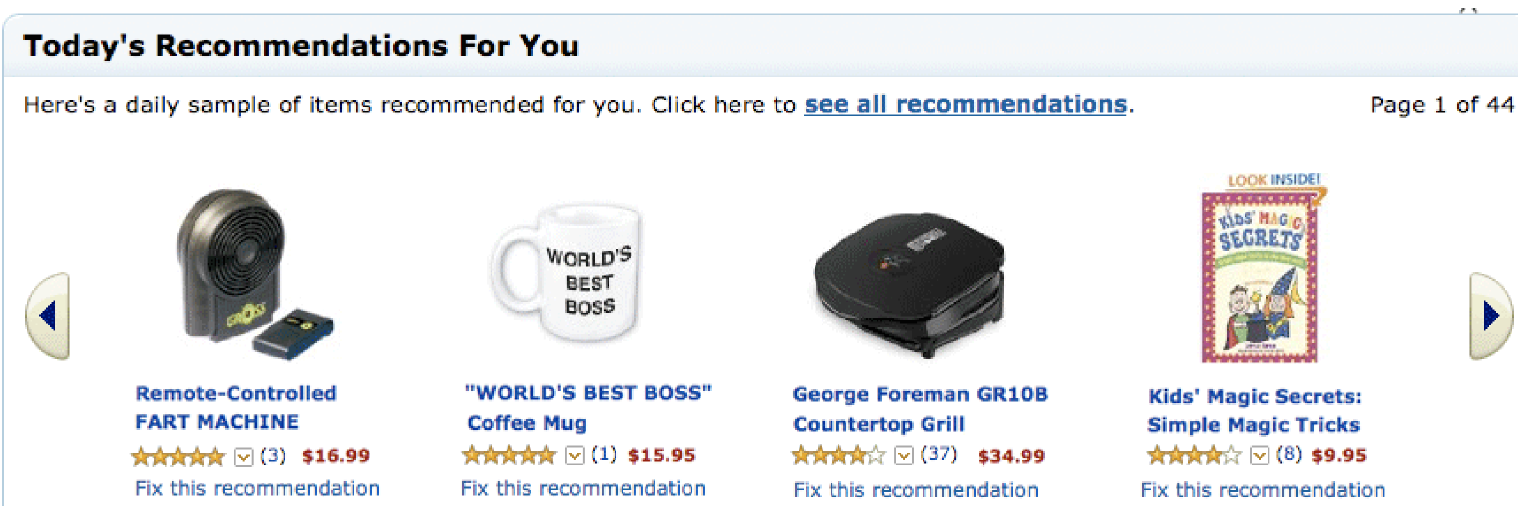
What is Data Science?

- relatively **new field** due to recent boost in digital transforation $\rightarrow$ *digitization, IoT, paperless office, ...*
- needs mathematics, computer science, and domain knowledge
- well-known examples for successfully harvesting and interpreting data:
  - $\rightarrow$ Google
  - $\rightarrow$ Spotify
  - $\rightarrow$ Amazon



What is Data Science?

- relatively **new field** due to recent boost in digital transforation $\rightarrow$ *digitization, IoT, paperless office, ...*
- needs mathematics, computer science, and domain knowledge
- well-known examples for successfully harvesting and interpreting data:
  - $\rightarrow$ Google
  - $\rightarrow$ Spotify
  - $\rightarrow$ Amazon



Data Science Cycle

iteratively...

1. extract and process / correct data
2. extract hypotheses
3. *develop approaches, models, algorithms, implementations (from statistics, optimization, numerics and simulation, math theory, databases, AI,...)*
4. use approaches to explore and understand data
5. derive predictions, decisions, consequences, recommendations for application domain
6. visualize data and results, possibly iterate

Your study programme covers these aspects.

Data Science Cycle

Goals of this lecture

- learn fundamental math-based data science concepts and algorithms
- understanding the underlying mathematical reasons why certain algorithms work well and others don’t
- solve realistic problems

Data Science Cycle

Goals of this lecture

- learn fundamental math-based data science concepts and algorithms
- understanding the underlying mathematical reasons why certain algorithms work well and others don't
- solve realistic problems

topics: unsupervised learning and supervised learning methods, e.g.

- clustering
- PCA, kernel methods, kernel-PCA
- statistical methods
- machine learning via neural networks
- graphical models,...

Some Data Science Examples

- prediction of solar injection in energy networks and best-possible curtailment decisions beforehand to avoid breakdown of the network
- learn customer / market preferences, produce accordingly and forecast price developments
- forecast development of chronic diseases, determine best possible medication and treatment
- forecast development of Covic-19, determine best possible reactions
- many more

Further Reading

- Hastie, Tibshirani, Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer
- www.deeplearningbook.org

Rough Differentiation in Learning Methods

supervised learning:

- predict values of an outcome measure based on a number of input measures (e.g., given some patient data together with label 'has illness' / 'does not have illness'. New patient data comes in, predict whether s/he is ill or not.)

Rough Differentiation in Learning Methods

supervised learning:

- predict values of an outcome measure based on a number of input measures (e.g., given some patient data together with label ’has illness’ / ’does not have illness’. New patient data comes in, predict whether s/he is ill or not.)

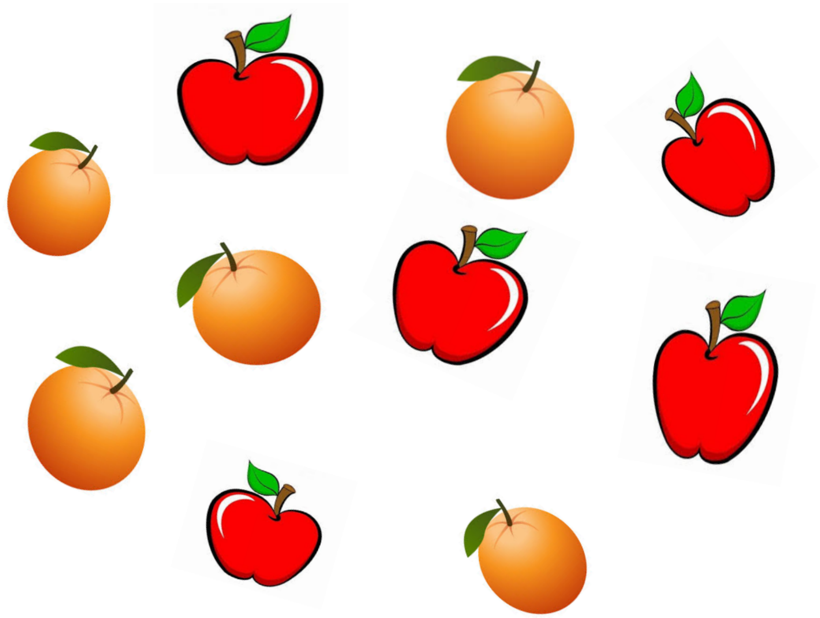
unsupervised learning:

- no outcome measure given. goal: find structures among data.

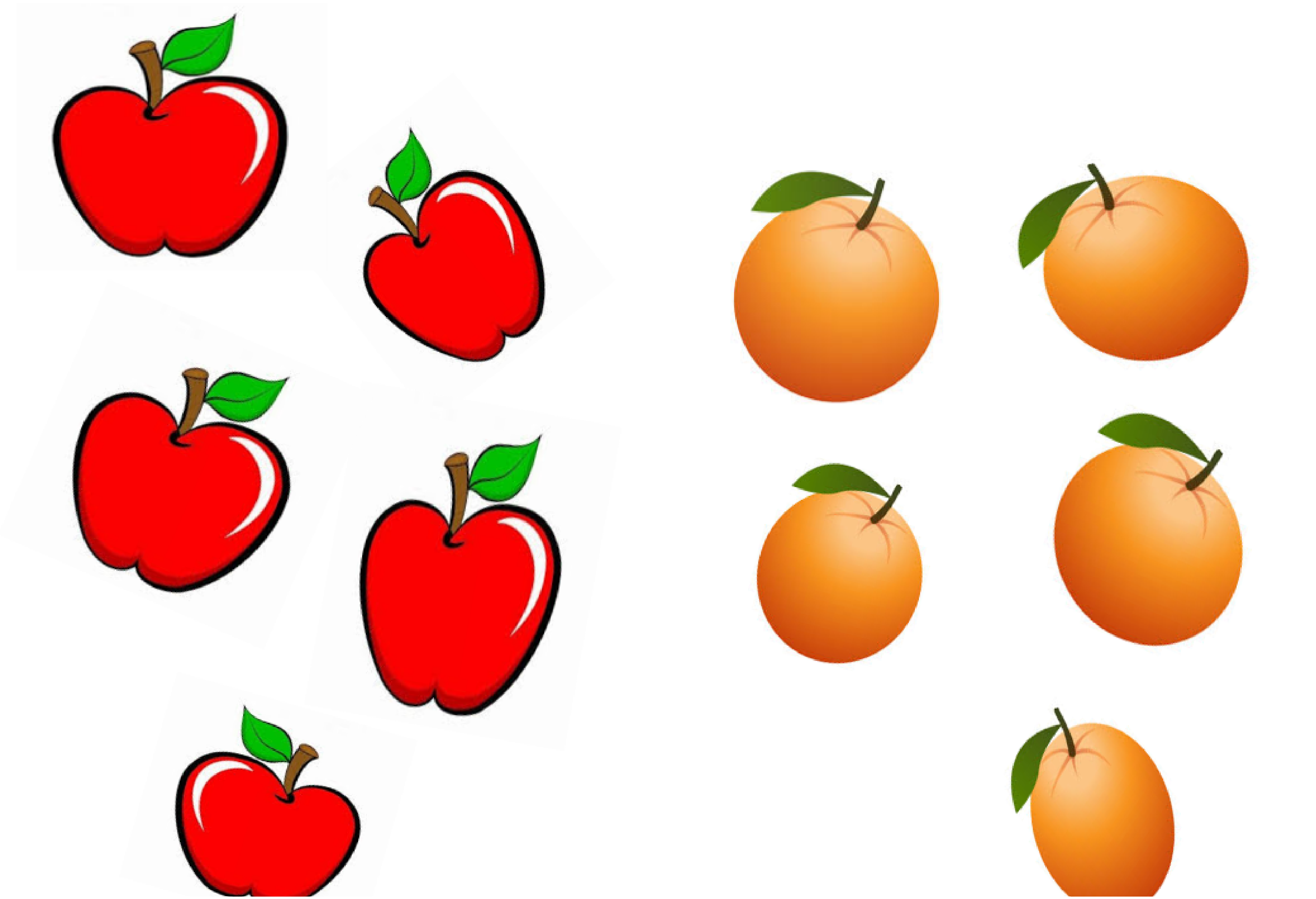
...also something in between: semi-supervised learning.

Unsupervised Learning: Clustering of Data

Is there any structure in this data?



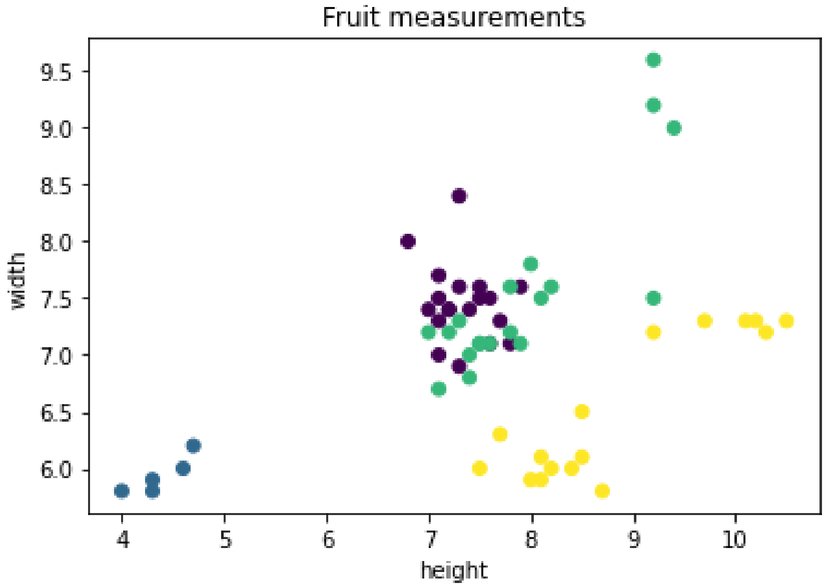
There are **two categories/clusters** such that objects *within* a cluster resemble each other but objects from *different clusters* look different.



What's clustering?

Given fruit measurement data $(\text{height}_i, \text{width}_i)_{i=1}^{N}$.

- Visually: There are multiple "categories" of data.



What's clustering?

Given fruit measurement data $(\text{height}_i, \text{width}_i)_{i=1}^{N}$.

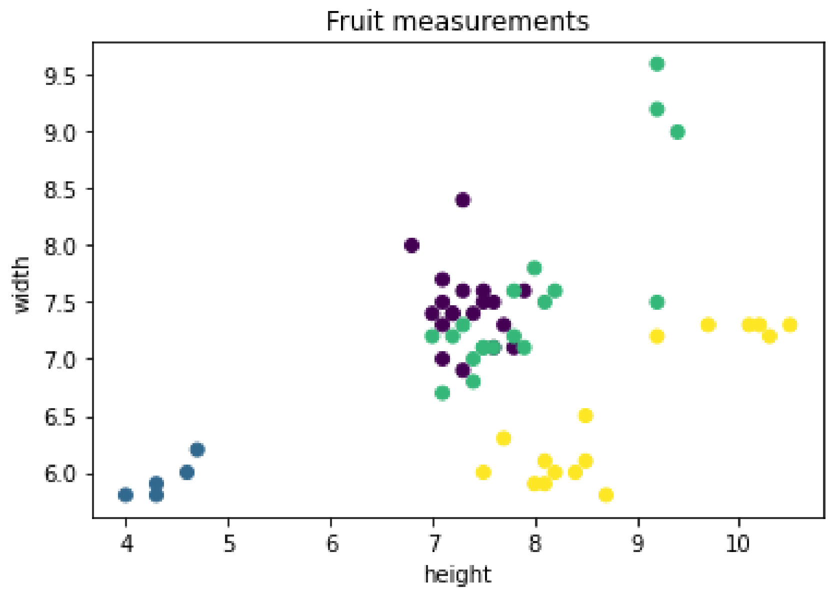
- Visually: There are multiple "categories" of data.
- How can we sort data into categories/clusters?

What's clustering?

Given fruit measurement data $(\text{height}_i, \text{width}_i)_{i=1}^N$.

- Visually: There are multiple "categories" of data.
- How can we sort data into categories/clusters?

$\rightarrow$ Clustering problem



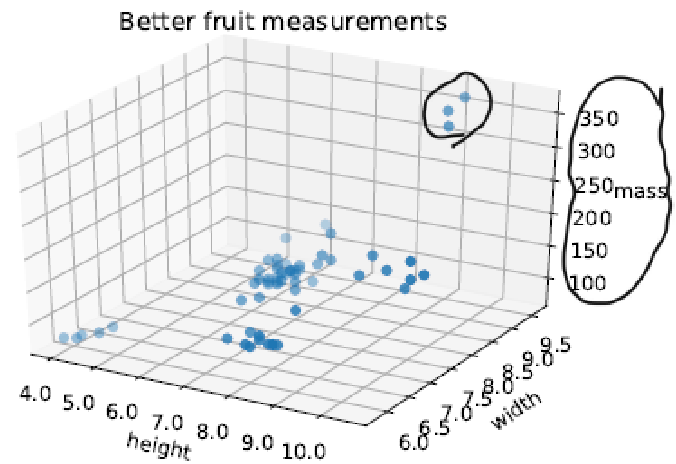
What's clustering?

Given fruit measurement data $(\text{height}_i, \text{width}_i)_{i=1}^{N}$.

- Visually: There are multiple "categories" of data.
- How can we sort data into categories/clusters?

$\rightarrow$ Clustering problem

- More measurements help



What’s clustering?

Given fruit measurement data $(\text{height}_i, \text{width}_i)_{i=1}^N$.

- Visually: There are multiple “categories” of data.
- How can we sort data into categories/clusters?

→ Clustering problem

- More measurements help
- Clustering the raw data by hand is cumbersome

	mass	width	height	color_score
0	192	8.4	7.3	0.55
1	180	8.0	6.8	0.59
2	176	7.4	7.2	0.60
3	86	6.2	4.7	0.80
4	84	6.0	4.6	0.79
5	80	5.8	4.3	0.77
6	80	5.9	4.3	0.81
7	76	5.8	4.0	0.81
8	178	7.1	7.8	0.92
9	172	7.4	7.0	0.89
10	166	6.9	7.3	0.93
11	172	7.1	7.6	0.92
12	154	7.0	7.1	0.88

Clustering Algorithm

Given

- N number of data points
- M number of variables (i.e "mass", "price", "color", ...)
- Data $X = \{x_1, \ldots, x_N\}$, where $x_n \in \mathbb{R}^M$ for all $n = 1, \ldots, N$
- K number of *assumed* clusters

Want

Clustering Algorithm

Given

- N number of data points
- M number of variables (i.e “mass”, “price”, “color”, ...)
- Data $X = \{x_1, \ldots, x_N\}$, where $x_n \in \mathbb{R}^M$ for all $n = 1, \ldots, N$
- K number of *assumed* clusters

Want

- *Assignment:* $x_n \mapsto k_n \in \{1, \ldots, K\}$ for all $n = 1, \ldots, N$

Clustering Algorithm

Given

- N number of data points
- M number of variables (i.e "mass", "price", "color", ...)
- Data $X = \{x_1, \dots, x_N\}$, where $x_n \in \mathbb{R}^M$ for all $n = 1, \dots, N$
- K number of *assumed* clusters

Want

- *Assignment:* $x_n \mapsto k_n \in \{1, \dots, K\}$ for all $n = 1, \dots, N$
- *Assignment rule:* $\mathbf{x} \mapsto k(\mathbf{x}) \in \{1, \dots, K\}$ for all $x \in \mathbb{R}^M$

Clustering Algorithm

Given

- N number of data points
- M number of variables (i.e “mass”, “price”, “color”, ...)
- Data $X = \{x_1, \ldots, x_N\}$, where $x_n \in \mathbb{R}^M$ for all $n = 1, \ldots, N$
- K number of *assumed* clusters

Want

- *Assignment:* $x_n \mapsto k_n \in \{1, \ldots, K\}$ for all $n = 1, \ldots, N$
- *Assignment rule:* $\mathbf{x} \mapsto k(\mathbf{x}) \in \{1, \ldots, K\}$ for all $x \in \mathbb{R}^M$
- *Reconstruction rule (’representative’):* $k \mapsto m_k \in \mathbb{R}^M$

On an abstract level:

- Determination of best possible clustering (w.r.t. some objective) is a classical combinatorial optimization problem
- K-means clustering: Determine k points, i.e., centers, that minimize the sum of the squared Eucidean distance to its closest center.

Clustering Algorithm

- Already in simplified / restricted situations, the problem is difficult, i.e., NP-hard. This means, we cannot expect to be able to determine an efficient algorithm that can efficiently determine the best clustering within polynomial time in the input size.
- more specifically: M. Mahajan, P. Nimbhorkar, K. Varadarajan: The planar k-means problem is NP-hard. Proceedings of WALCOM: Algorithms and Computation, S.274-285 (2009)

K-means clustering as optimization problem

Find clustering $\underline{C} = \{C_1, \dots, C_K\}$ into sets $C_k \subset X$ and centers $\underline{m} = \{m_1, \dots, m_k\}$ with $m_k \in C_k$, which minimize the clustering energy

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

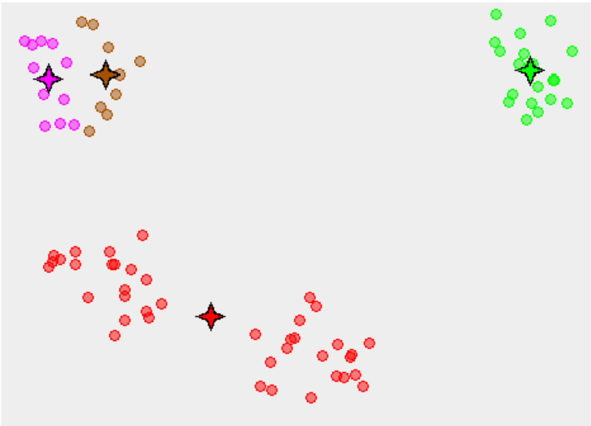
K-means clustering as optimization problem

Find clustering $\underline{C} = \{C_1, \ldots, C_K\}$ into sets $C_k \subset X$ and centers $\underline{m} = \{m_1, \ldots, m_k\}$ with $m_k \in C_k$, which minimize the clustering energy

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

Observations

- The clustering energy has local minima[1]



Derivation of the K-means algorithm

Let us fix the clustering $\underline{C}$ in

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 \,.$$

Derivation of the K-means algorithm

Let us fix the clustering $\underline{C}$ in

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

necessary first-order optimality condition: gradient with respect to m_k is zero, i.e., a critical point.
Taking the gradient with respect to m_k we obtain the first-order optimality condition:

$$0 = \nabla_{m_k} E(\underline{C}, \underline{m}) = \sum_{x \in C_k} (x - m_k) = \sum_{x \in C_k} x - |C_k| m_k$$

Derivation of the K-means algorithm

Let us fix the clustering $\underline{C}$ in

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

necessary first-order optimality condition: gradient with respect to m_k is zero, i.e., a critical point.
Taking the gradient with respect to m_k we obtain the first-order optimality condition:

$$0 = \nabla_{m_k} E(\underline{C}, \underline{m}) = \sum_{x \in C_k} (x - m_k) = \sum_{x \in C_k} x - |C_k| m_k$$

and hence

$$m_k = \frac{1}{|C_k|} \sum_{x \in C_k} x \mathrel{\widehat{=}} \text{mean of the cluster}$$

problem: do not know the means, thus heuristically search for good means

Derivation of the K-means algorithm

Conversely, let us fix the means $\underline{m}$ in

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

Derivation of the K-means algorithm

Conversely, let us fix the means $\underline{m}$ in

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

perform the simple assignment step

$$C_k = \{x \in X \, : \, \|x - m_k\| \leq \left\|x - m_j\right\| \text{ for all } j = 1, \ldots, K\}$$

$\widehat{=}$ Voronoi cell of m_k

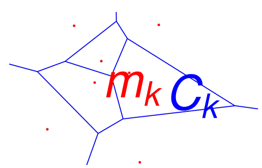
Derivation of the K-means algorithm

Conversely, let us fix the means $\underline{m}$ in

$$E(\underline{C}, \underline{m}) := \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2 .$$

perform the simple assignment step

$$C_k = \{x \in X \,:\, \|x - m_k\| \leq \left\|x - m_j\right\| \text{ for all } j = 1, \ldots, K\}$$
$$\widehat{=} \text{ Voronoi cell of } m_k$$



K-means clustering

Data: $X = \{x_1, \ldots, x_N\}$ and number of clusters $K \in \mathbb{N}$
Result: cluster means $\underline{m} = (m_1, \ldots, m_K)$

---

initialize $\underline{m}$ randomly;
repeat
  `// assignment step:`
  for $n \leftarrow 1$ **to** N *// assign n-th point to cluster with nearest mean* **do**
    $k_n \leftarrow \operatorname{argmin}_k \|x_n - m_k\|$
  end
  `// update step:`
  for $k \leftarrow 1$ **to** K **do**
    $C_k \leftarrow \{n \in \{1, \ldots, N\} : k_n = k\}$; `// cluster`
    $m_k \leftarrow \frac{1}{|C_k|} \sum_{n \in C_k} x_n$; `// mean of current cluster`
  end
until *assignment step does not do anything*;

K-means clustering

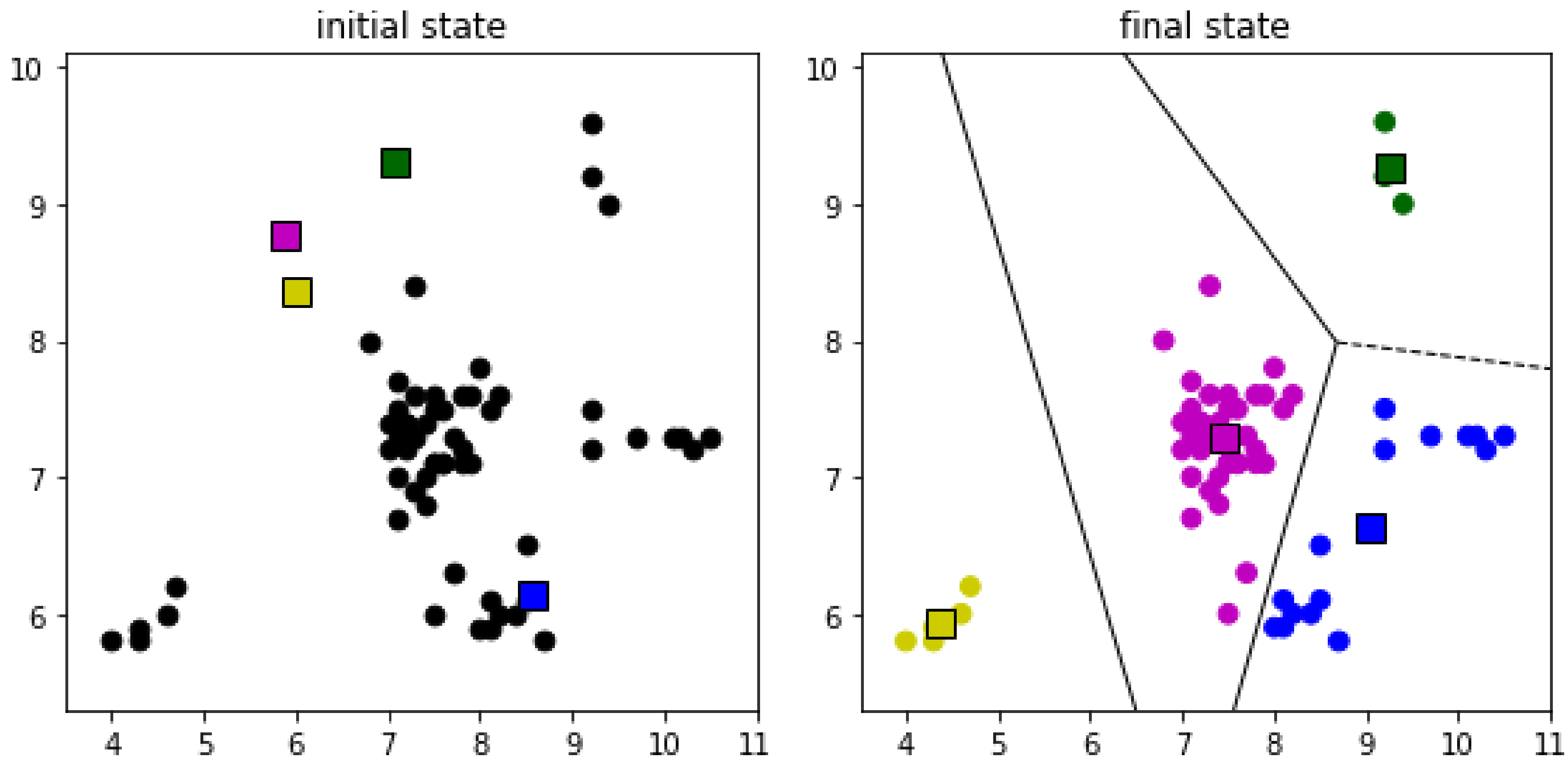
Data: $X = \{x_1, \dots, x_N\}$ and number of clusters $K \in \mathbb{N}$
Result: cluster means $\underline{m} = (m_1, \dots, m_K)$

---

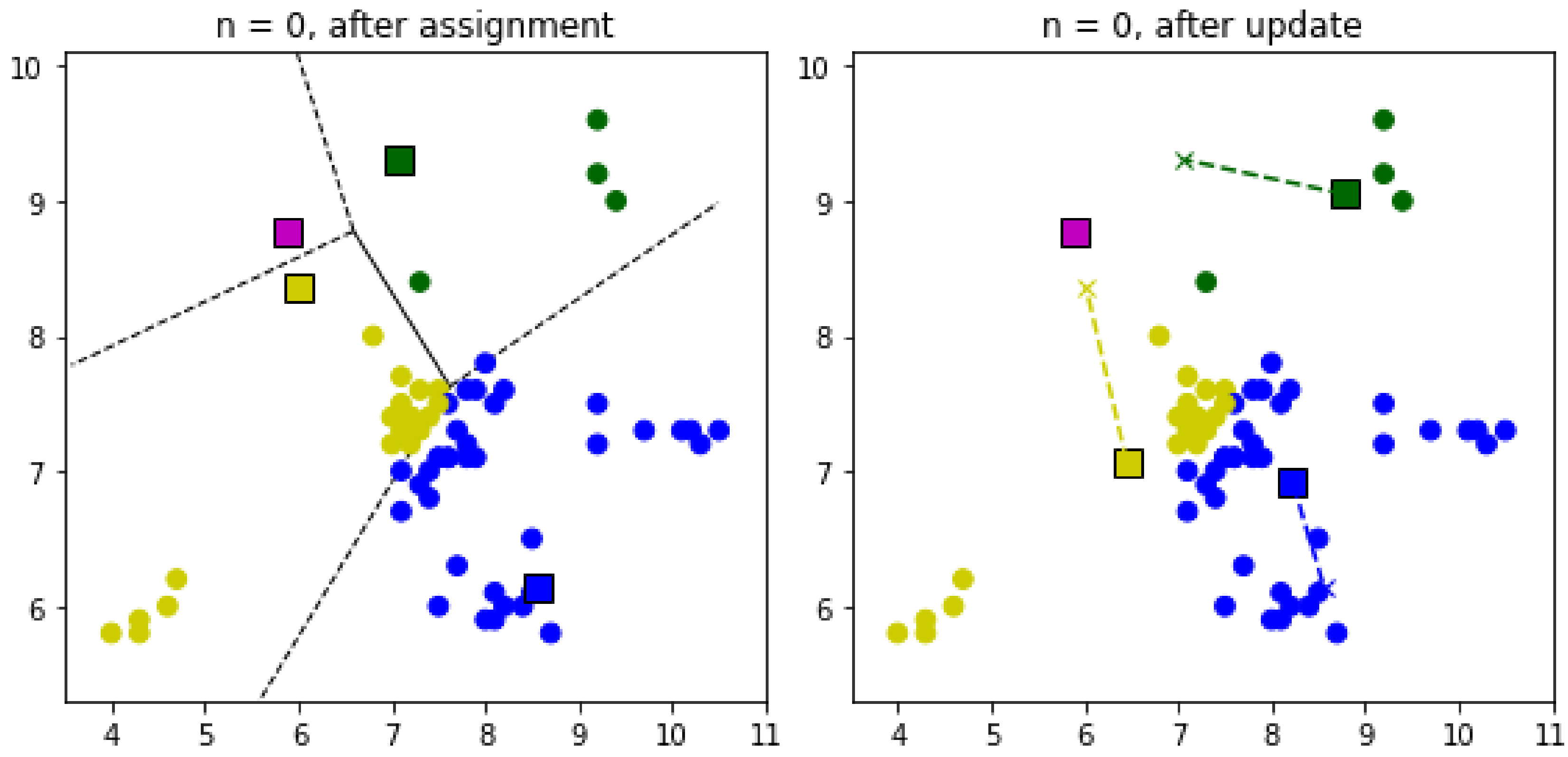
initialize $\underline{m}$ randomly;
repeat
    `// assignment step:`
    for $n \leftarrow 1$ **to** N *// assign n-th point to cluster with nearest mean* **do**
        $k_n \leftarrow \operatorname{argmin}_k \|x_n - m_k\|$
    end
    `// update step:`
    for $k \leftarrow 1$ **to** K **do**
        $C_k \leftarrow \{n \in \{1, \dots, N\} : k_n = k\}$; `// cluster`
        $m_k \leftarrow \frac{1}{|C_k|} \sum_{n \in C_k} x_n$; `//mean of current cluster`
    end
until *assignment step does not do anything*;

- Assignment rule: $\mathbf{x} \mapsto \operatorname{argmin}_k \|x - m_k\|$.
- Reconstruction rule: $k \mapsto m_k$

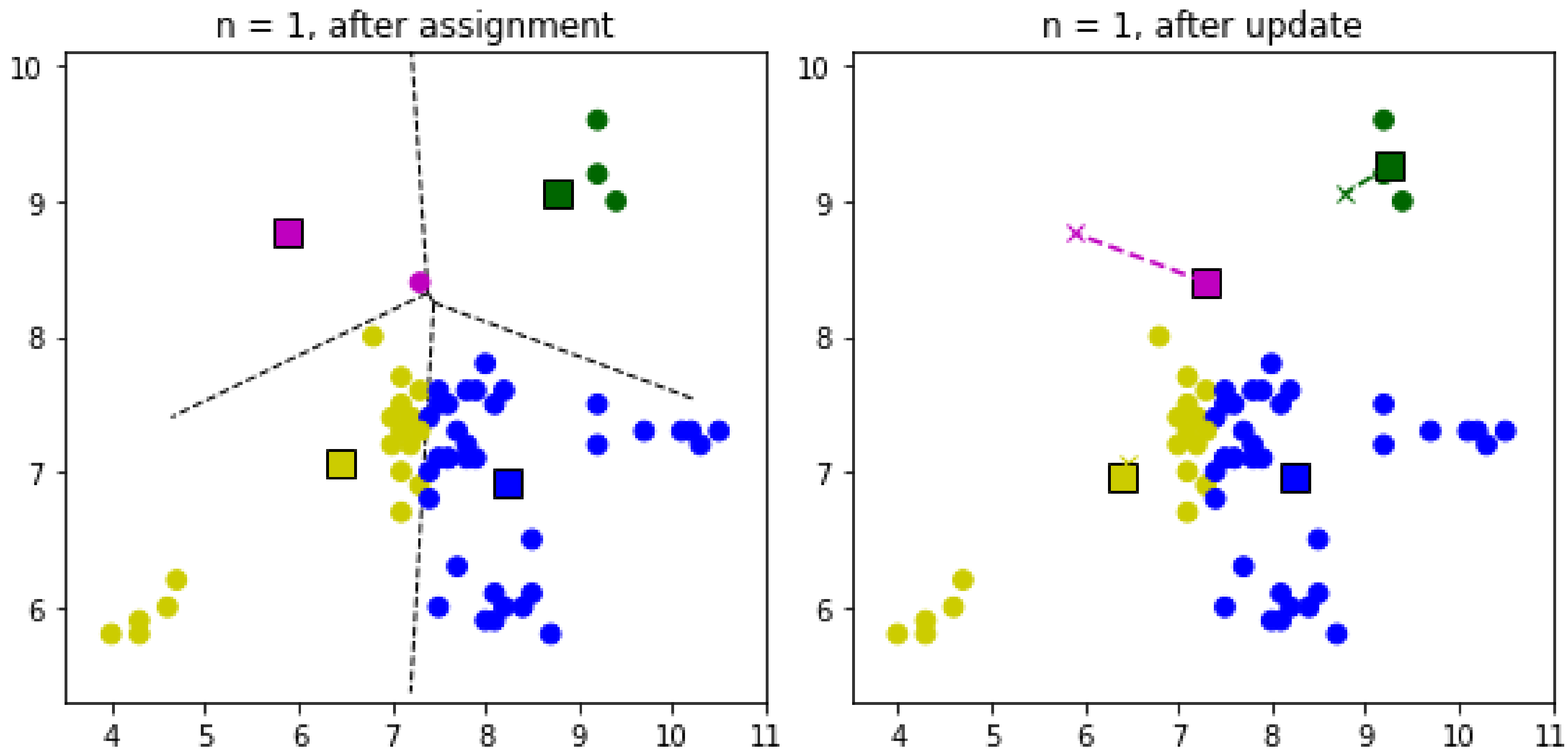
Back to our fruits data set



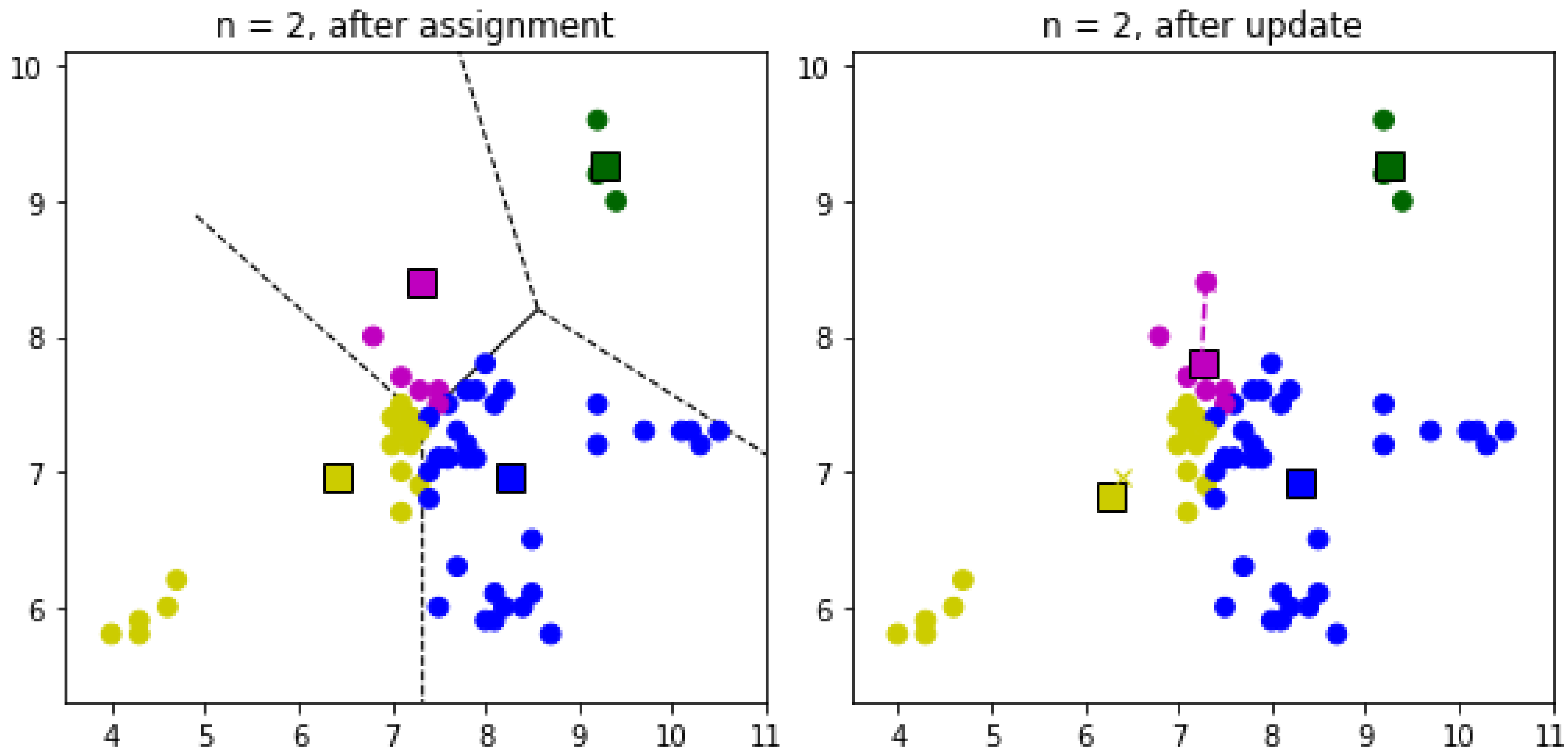
Back to our fruits data set



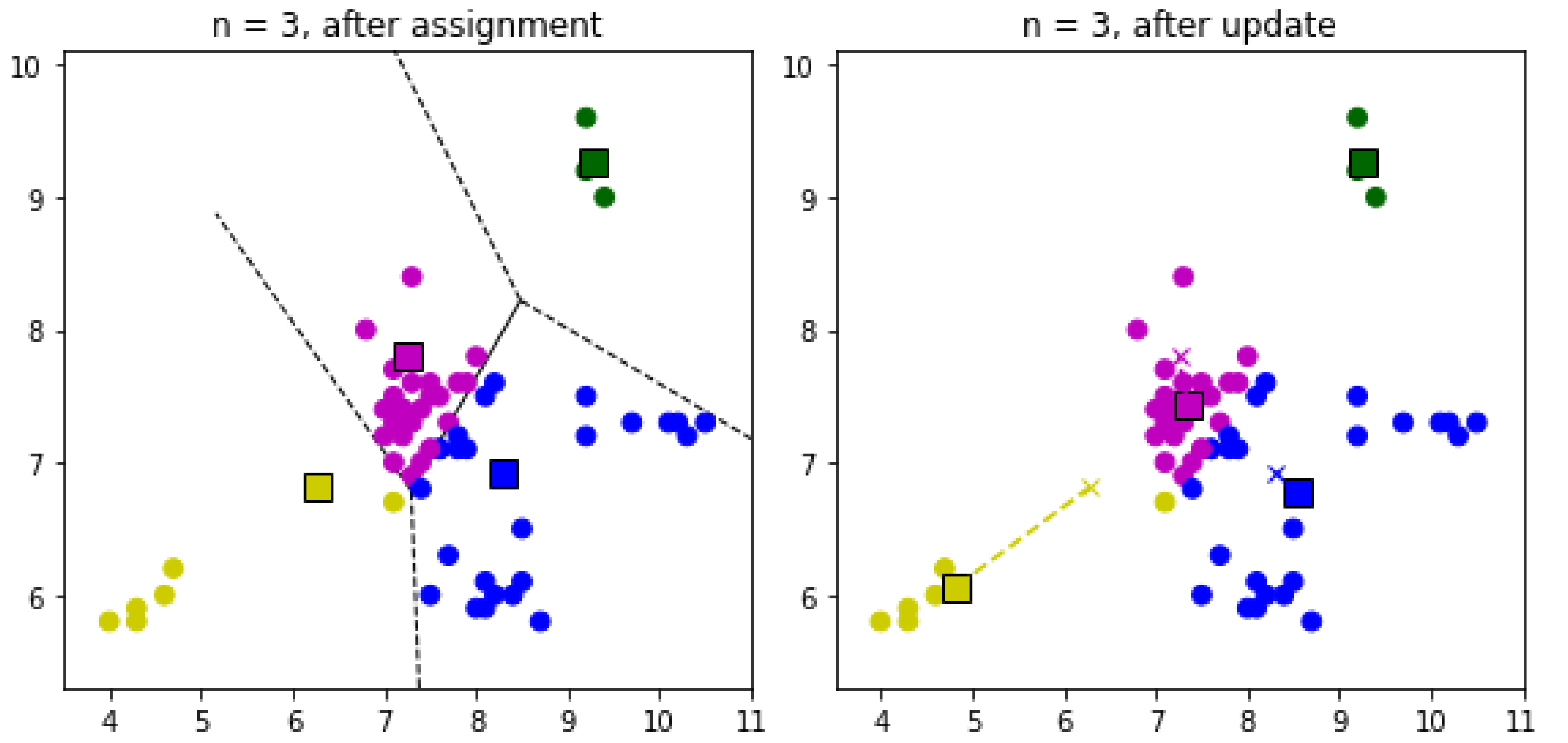
Back to our fruits data set



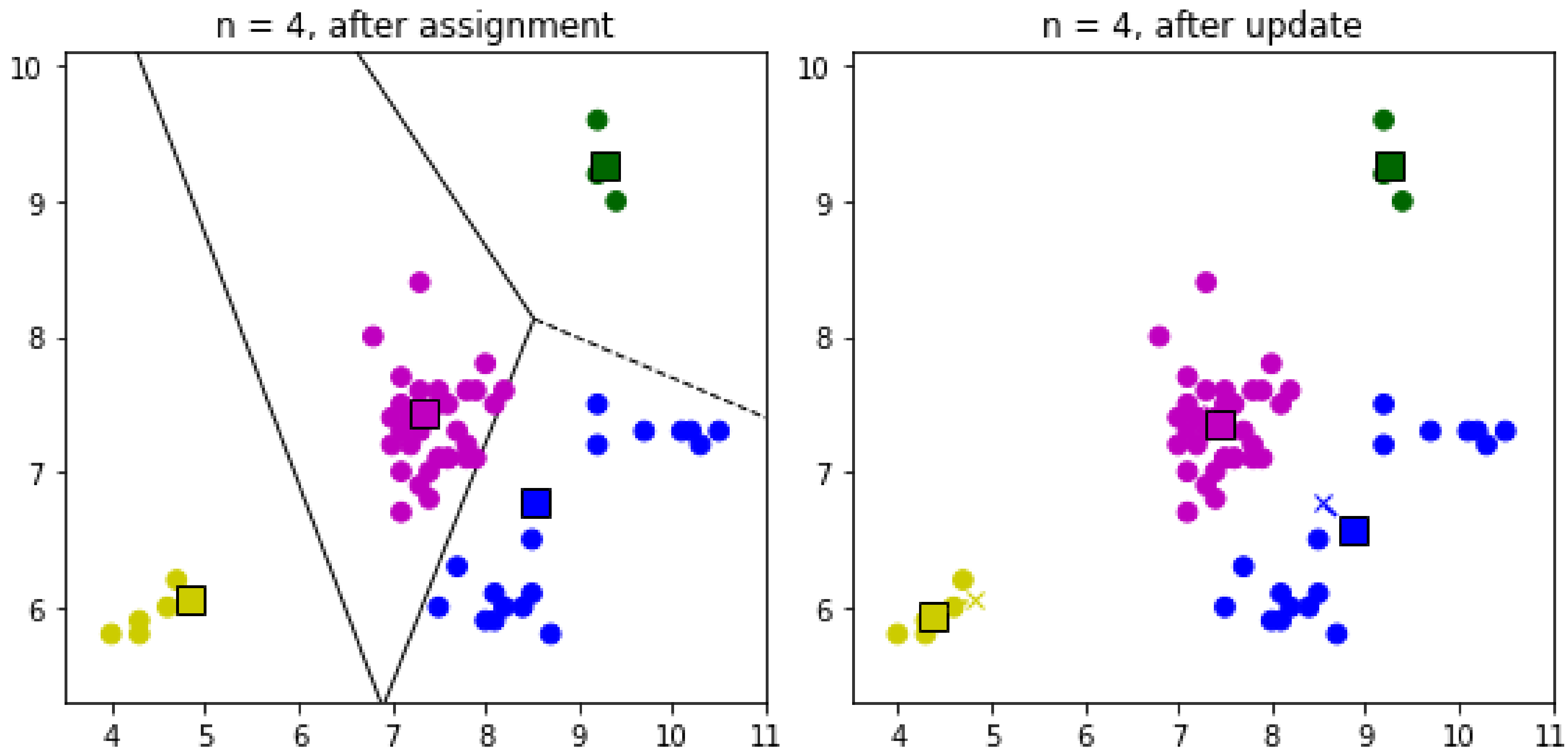
Back to our fruits data set



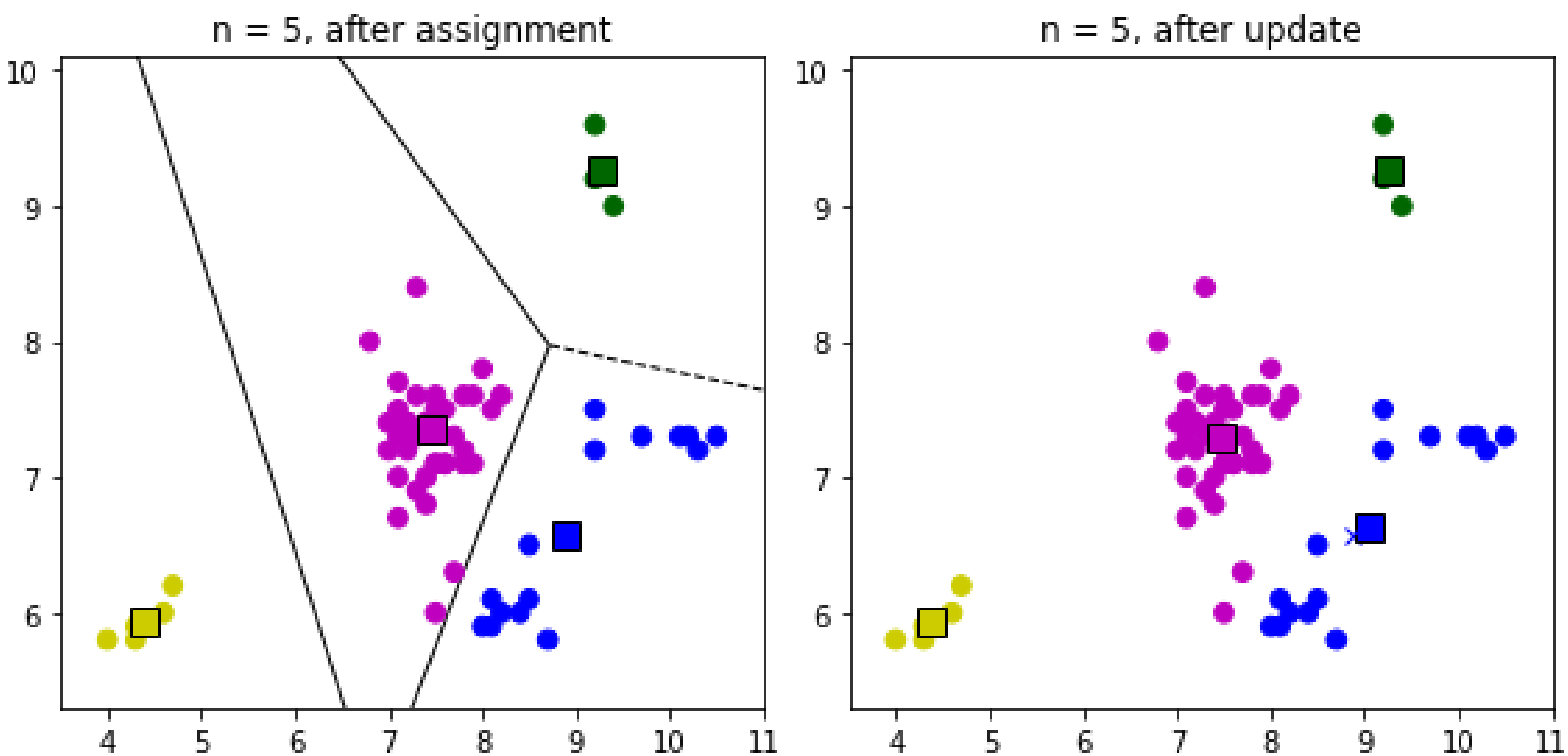
Back to our fruits data set



Back to our fruits data set

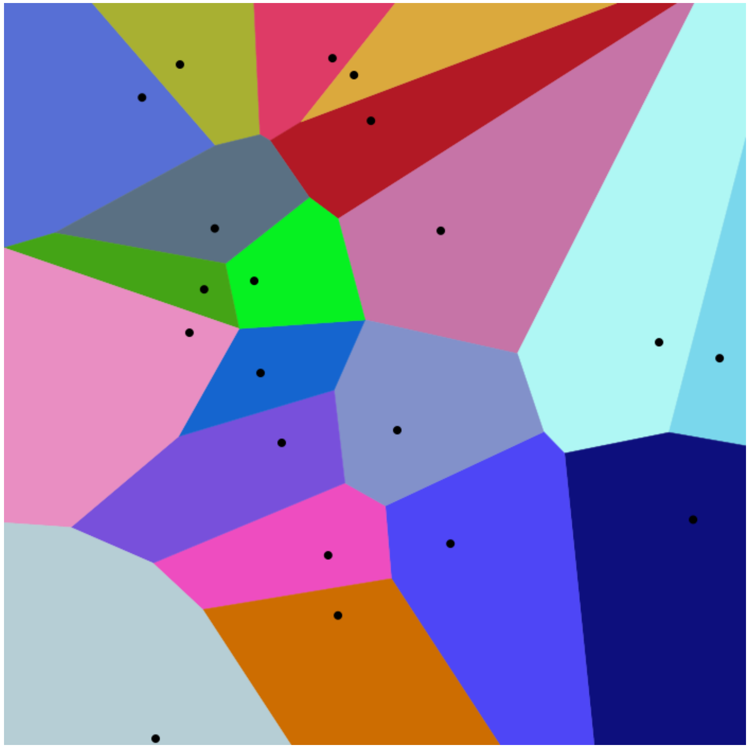


Back to our fruits data set



K-means

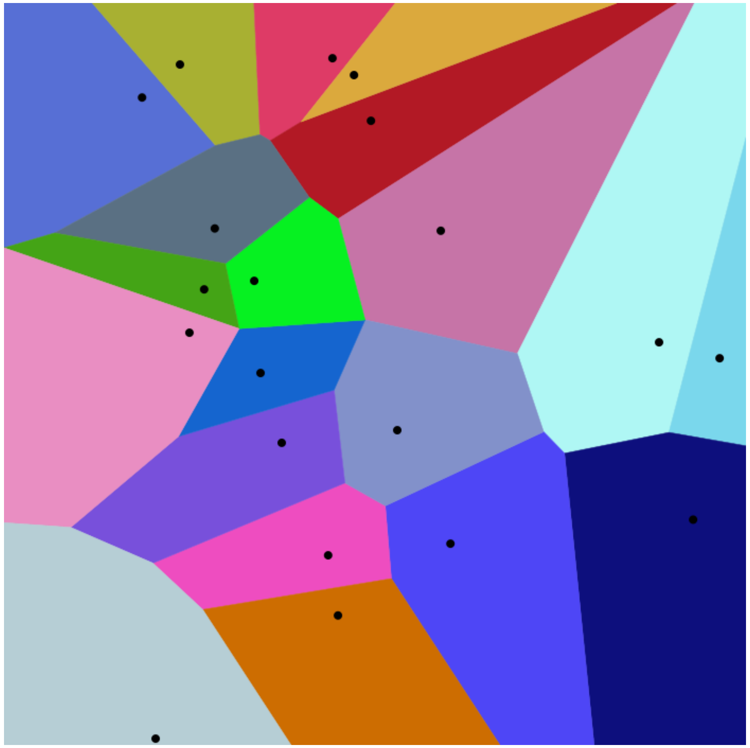
Euclidean Voronoi cells induce round clusters and straight interfaces



$$E(\underline{C}, \underline{m}) = \frac{1}{2} \sum_{k=1}^{K} \sum_{x \in C_k} \|x - m_k\|^2$$

K-means

Euclidean Voronoi cells induce round clusters and straight interfaces



$$E(\underline{C}, \underline{m}) = \frac{1}{2}\sum_{k=1}^{K}\sum_{x\in C_k}\|x - m_k\|^2$$

$$E(\underline{C}, \underline{m}) = \sum_{k=1}^{K}\sum_{x\in C_k}\|x - m_k\|_1$$

Advantages and Disadvantages of K-Means

advantages

- easy to implement
- can run with only a set of real data vectors and value of k
- feasible clustering is always available

Advantages and Disadvantages of K-Means

advantages

- easy to implement
- can run with only a set of real data vectors and value of k
- feasible clustering is always available

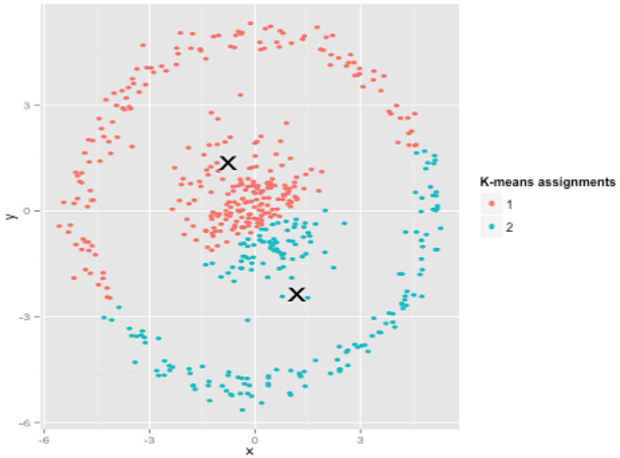
disadvantages

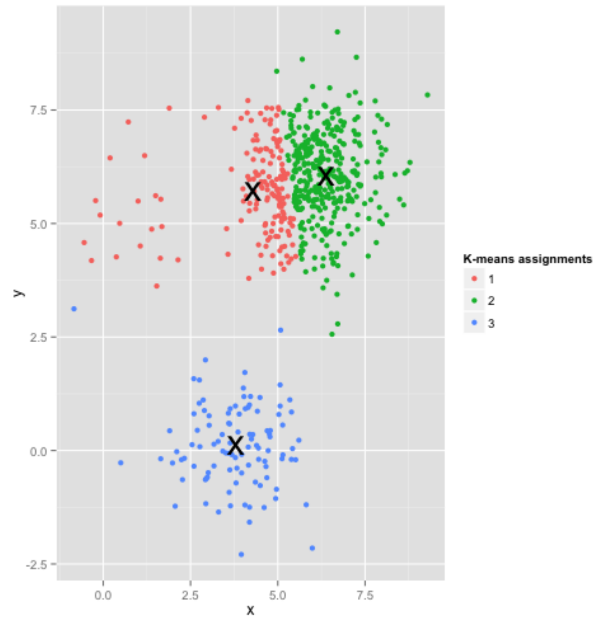
- Choosing a good value of k can be difficult. Typically, test several values
- assumes numerical data, not categorial data ('car', 'truck'...)
- K-means aims at minimizing the Euclidean distances. This is not always the right objective.
- result strongly depends on initialization (improvements known)
- assumes that clusters are convex.

Advantages and Disadvantages of K-Means

disadvantages:

- K-means sometimes does not work well. Can behave badly in non-spherical / nonconvex data or for unevenly sized clusters, i.e., it has some implicit assumptions





from varianceexplained.org

next: some improvements.