*** Applied Machine Learning Fundamentals *** Clustering

M. Sc. Daniel Wehner

SAP SE

Winter term 2019/2020

Find all slides on GitHub

Lecture Overview

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
Unit III	Bayesian Decision Theory
Unit IV	Probability Density Estimation
Unit V	Regression
Unit VI	Classification I
Unit VII	Evaluation
Unit VIII	Classification II
Unit IX	**Clustering**
Unit X	Dimensionality Reduction

Agenda for this Unit

Section:

Introduction

DHBW
Duale Hochschule
Baden-Württemberg

Clustering Introduction

- Clustering belongs to the category of **unsupervised learning**
- A clustering algorithm tries to **find structure** in the data
- Once the clusters are found, they first have to be interpreted...
- ...and can then be used for prediction purposes

A cluster must be **internally homogeneous**, but simultaneously **externally heterogeneous**. (Elements of one cluster have to be very similar, but must differ significantly from elements in other clusters.)

Example Use Cases for Clustering

- **Behavioral segmentation**
  - Customer segmentation (e. g. **sinus milieus**)
  - Creating profiles based on activity monitoring
- **Sorting sensor measurements**
  - Image grouping
  - Detection of activity types in motion sensors
- **Inventory categorization**
  - Grouping inventory by sales activity
  - Grouping inventory by manufacturing metrics
- Many, many more, ...

Clustering Strategies

1. EM-based clustering, e. g.: k-Means
2. Hierarchical clustering, e. g.:
   - Agglomerative clustering
   - Divisive clustering
3. Affinity-based clustering, e. g.:
   - Spectral clustering
   - DBSCAN

Section:

k-Means

Duale Hochschule Baden-Württemberg

k-Means: Procedure

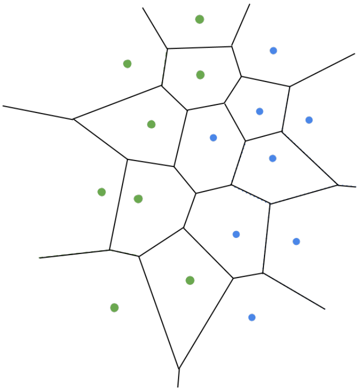
- The algorithm is an instance of vector quantization
  - It represent data points by a single vector (here: centroid) which is close to them
  - This is useful for **compression**!
- **How to**: Create k partitions ($\widehat{=}$ clusters) of the data set $\mathcal{D}$, such that the sum of squared deviations from the cluster centroids is **minimal**:

$$\min_{\boldsymbol{\mu}_j} \sum_{j=1}^{k} \sum_{\boldsymbol{x}^{(i)} \in \mathcal{D}_j} \left\| \boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j \right\|^2 \tag{1}$$

- With $\mathcal{D}_j \equiv j^{th}$ cluster, $\mu_j \equiv$ centroid of j^{th} cluster

Result: Voronoi Diagram

- The dots represent cluster centroids
- The lines visualize the **cluster boundaries**
- For a new point we can easily determine to which cluster it has to be assigned



k-Means Algorithm

- Input: $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \dots, \boldsymbol{x}^{(n)}\} \in \mathbb{R}^{n \times m}$, Number of clusters k
- Algorithm:
  1. $t \longleftarrow 1$
  2. Randomly choose k means $\boldsymbol{\mu}_1^{\langle 1 \rangle}, \boldsymbol{\mu}_2^{\langle 1 \rangle}, \dots, \boldsymbol{\mu}_k^{\langle 1 \rangle}$
  3. While not converged:

     3a Assign each $\boldsymbol{x}^{(i)} \in \mathcal{D}$ to the closest cluster:

$$\mathcal{D}_j^{\langle t \rangle} = \left\{ \boldsymbol{x}^{(i)} : \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j^{\langle t \rangle}\|^2 \leqslant \|\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{j^*}^{\langle t \rangle}\|^2;\ \forall j^* = 1, 2, \dots, k; \boldsymbol{x}^{(i)} \in \mathcal{D} \right\}$$

     3b Update cluster centroids $\boldsymbol{\mu}_j$:

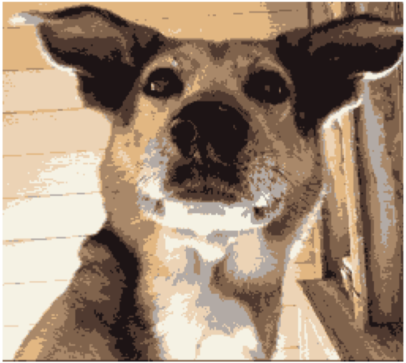
$$\boldsymbol{\mu}_j^{\langle t+1 \rangle} = \frac{1}{|\mathcal{D}_i^{\langle t \rangle}|} \sum_{\boldsymbol{x}^{(i)} \in \mathcal{D}_j^{\langle t \rangle}} \boldsymbol{x}^{(i)}$$

     3c $t \longleftarrow t + 1$

Image Compression

Original image

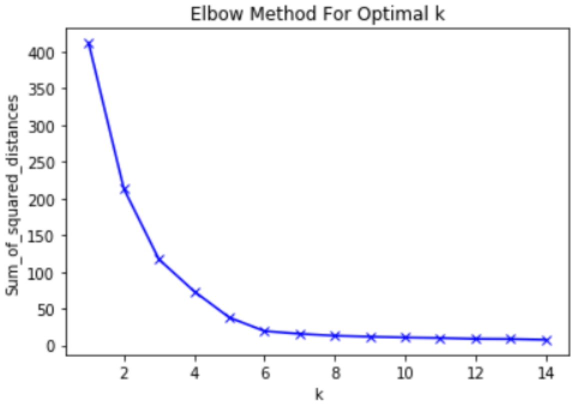
Compressed image



k-Means Issues

- The algorithm assumes all clusters are **sperical** ($\neq$ **affinity-based clustering**)
- Does not have a notion of **outliers** (unlike *DBSCAN*)
- What is the correct value for k? $\Rightarrow$ **Elbow-method:**
  - Measure sum of squred distances from data points to cluster centers (inertia)
  - Record results for different values for k and plot them
  - Search for the 'elbow point'

Elbow Method



Section:

Hierarchical Clustering

DHBW
Duale Hochschule
Baden-Württemberg

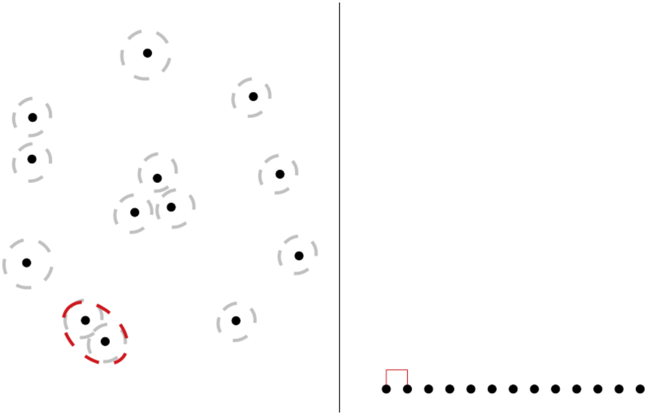
Agglomerative Clustering

1. Start with one cluster for each instance: $C = \{\{\boldsymbol{x}^{(i)}\} : \boldsymbol{x}^{(i)} \in \boldsymbol{X}\}$
2. Compute distance $d(C_i, C_j)$ between all pairs of clusters C_i, C_j
3. Join clusters C_i and C_j with minimum distance into a new cluster C_p:

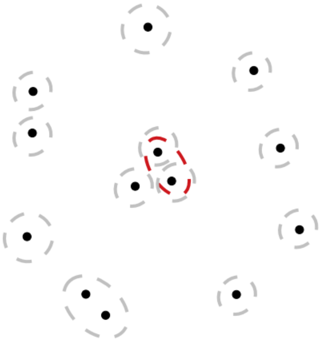
$$C_p = \{C_i, C_j\}$$
$$C = (C \setminus \{C_i, C_j\}) \cup \{C_p\}$$

4. Compute distances between C_p and all other clusters in C
5. If $|C| > 1$, goto 3

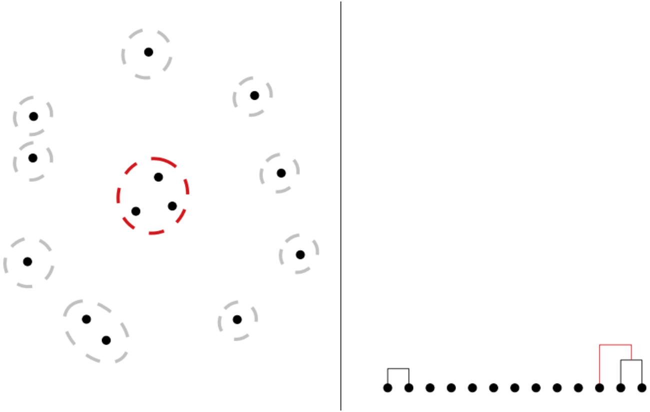
Agglomerative Clustering: Example



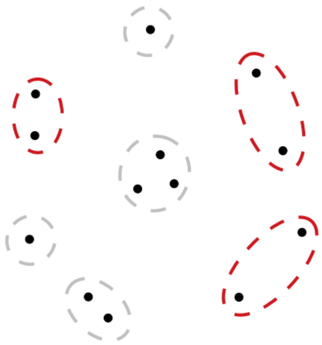
Agglomerative Clustering: Example (Ctd.)

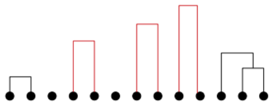


Agglomerative Clustering: Example (Ctd.)

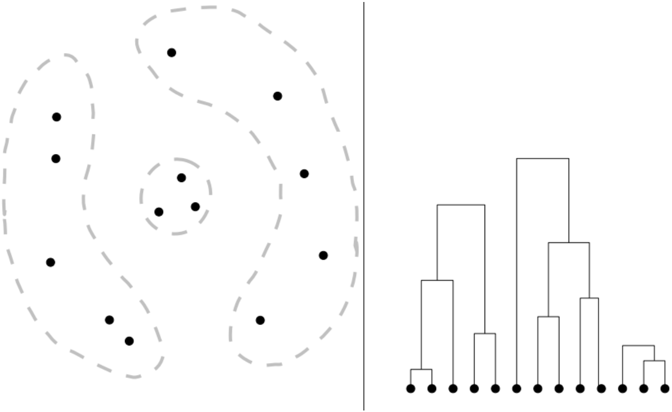


Agglomerative Clustering: Example (Ctd.)

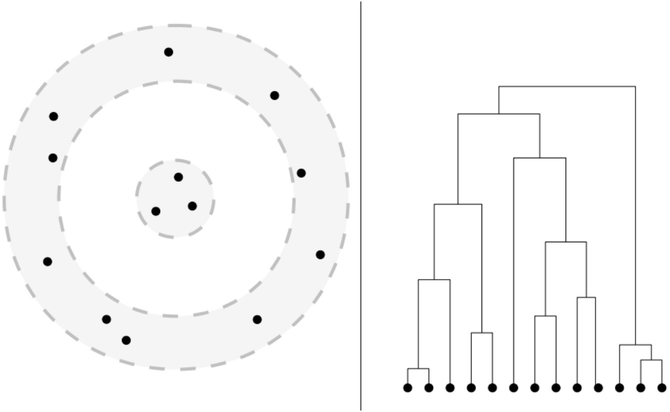




Agglomerative Clustering: Example (Ctd.)



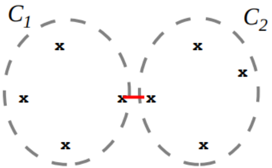
Agglomerative Clustering: Example (Ctd.)



Single Linkage

- Computing distances between clusters C_1 and C_2
- **Single linkage**:

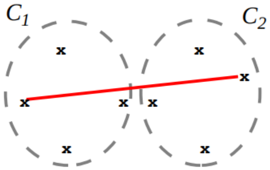
$$d(C_1, C_2) = \min\{d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) : \boldsymbol{x}^{(i)} \in C_1, \boldsymbol{x}^{(j)} \in C_2\}$$



Complete Linkage

- Computing distances between clusters C_1 and C_2
- **Complete linkage**:

$$d(C_1, C_2) = \max\{d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) : \boldsymbol{x}^{(i)} \in C_1, \boldsymbol{x}^{(j)} \in C_2\}$$



Section:

Spectral Clustering

DHBW
Duale Hochschule
Baden-Württemberg

Spectral Clustering

Section:

Wrap-Up

DHBW
Duale Hochschule
Baden-Württemberg

Summary

-

Self-Test Questions

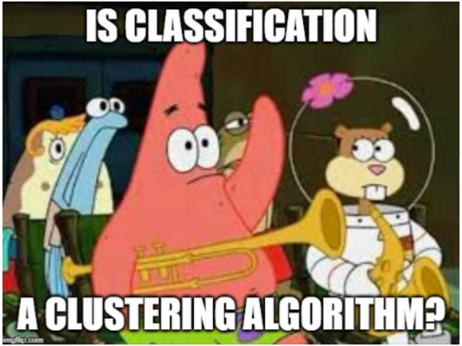
1.

What's next...?

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
Unit III	Bayesian Decision Theory
Unit IV	Probability Density Estimation
Unit V	Regression
Unit VI	Classification I
Unit VII	Evaluation
Unit VIII	Classification II
Unit IX	Clustering
Unit X	**Dimensionality Reduction**

Recommended Literature and further Reading

Meme of the Day



Thank you very much for the attention!

Topic: *** Applied Machine Learning Fundamentals *** Clustering
Term: Winter term 2019/2020

Contact:
M. Sc. Daniel Wehner
SAP SE
daniel.wehner@sap.com

Do you have any questions?