WeST People and Knowledge Networks | Institute for Web Science & Technologies

UNIVERSITÄT KOBLENZ · LANDAU

Machine Learning and Data Mining WS21/22

"4 Clustering I"

Dr. Zeyd Boukhers

@ZBoukhers

Institute for Web Science and Technologies

University of Koblenz-Landau

November 17, 2021

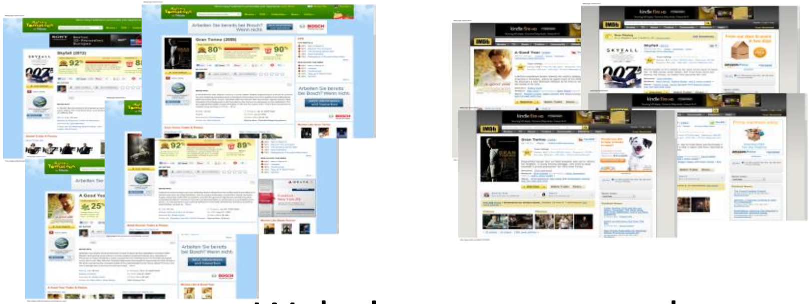
From previous lecture

- Data dimension reduction
    - PCA
    - SVD

Agenda

- What is clustering?
- What is unsupervised learning?
- How to evaluate clustering results?
- What are intrinsic and extrinsic evaluation measures?
- How does K-Means work?
- How to choose k for K-Means?
- What is the EM algorithm?

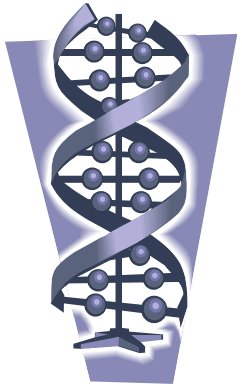
Examples



Web document templates

User Profiles

Documents

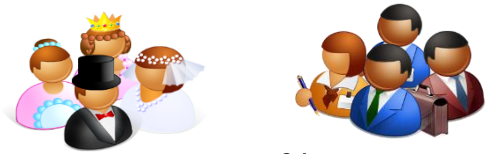


Genetic sequences

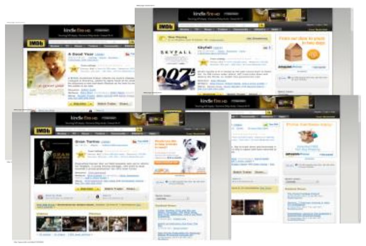
Language dialects

Goal of clustering

- Identification of a finite set of *clusters* (= categories, “classes“, groups) in the data.
- Objects in the same cluster should be as *similar* as possible.
  - High intra-similarity.
- Objects in different clusters should be as *dissimilar* as possible.
  - High inter-variance
- *“Unsupervised learning”* => no labels are given.

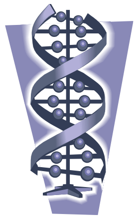


User profiles



Web document templates

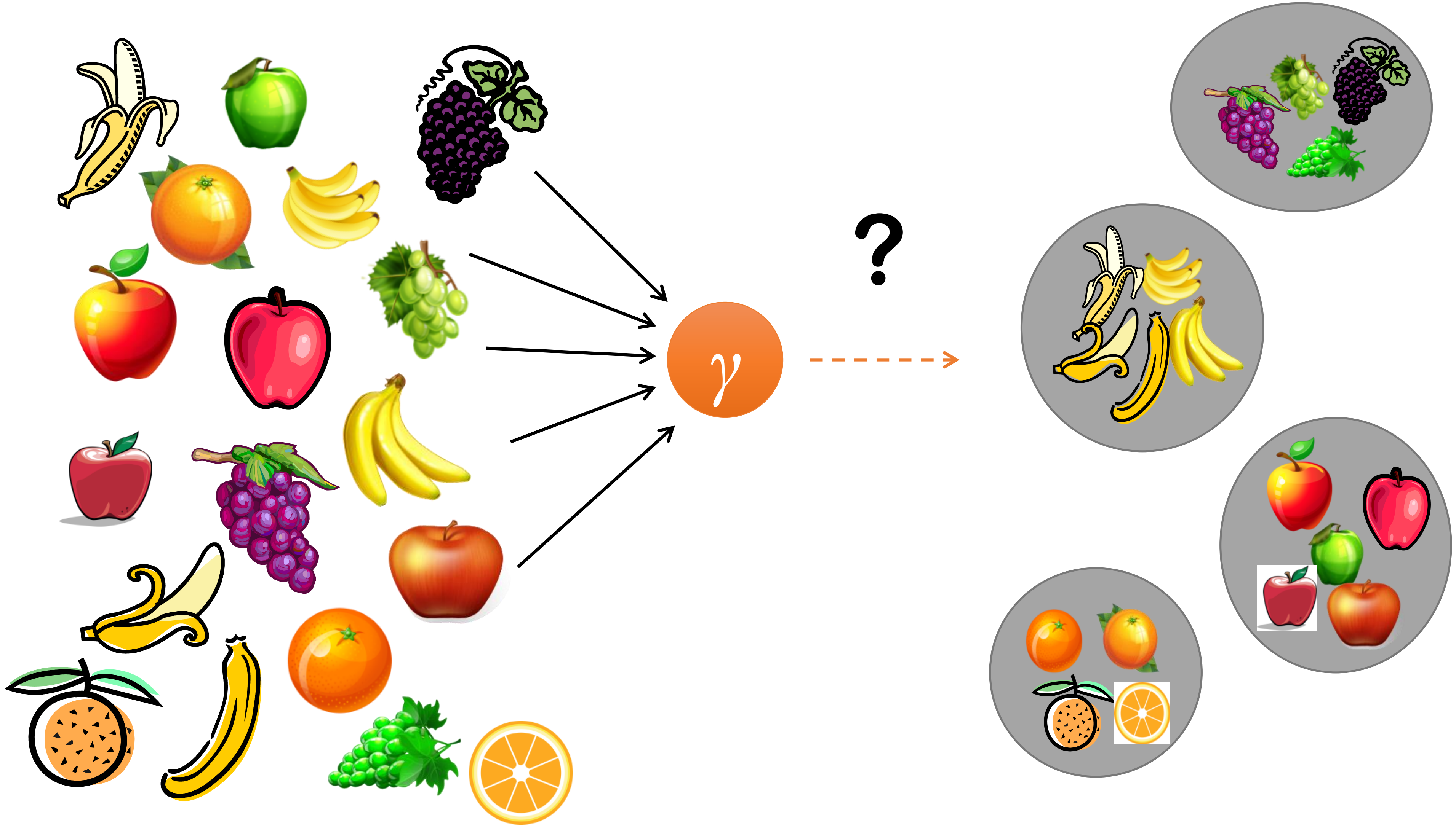
Documents



Genetic sequences

Language dialects

Clustering



What does similar mean?

What does similar mean?

Task

- Objects (dataset):
  - $D = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$

- An object is characterized by attributes
  - $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,l})$

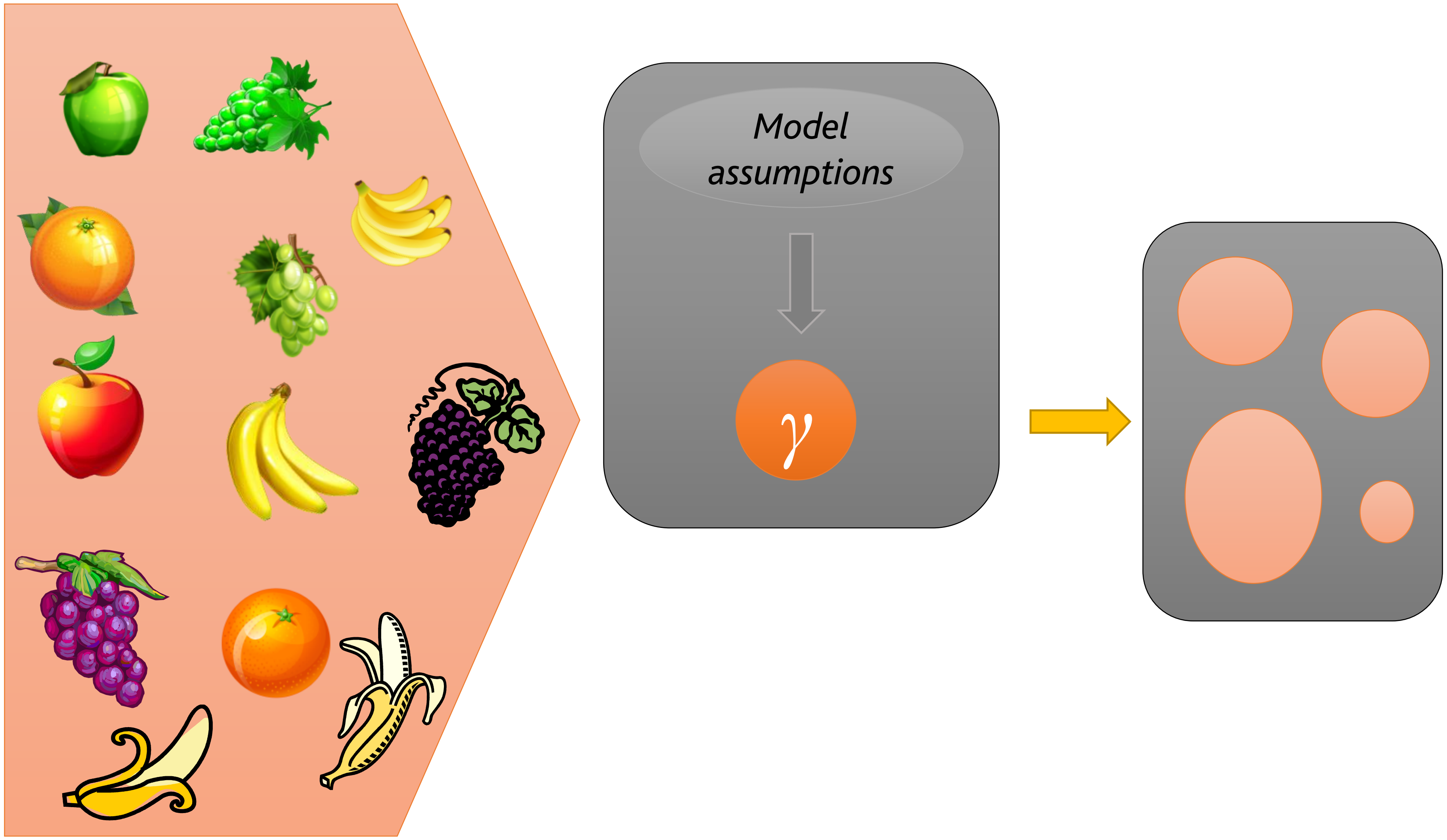
(green, round, even)

(orange, round, rough)

- Task:
  - Find groups
    - $\Psi = \{\psi_1, \psi_2, \dots, \psi_k\}$
  - Find function
    - $\gamma : D \rightarrow \Psi$
- Unsupervised learning

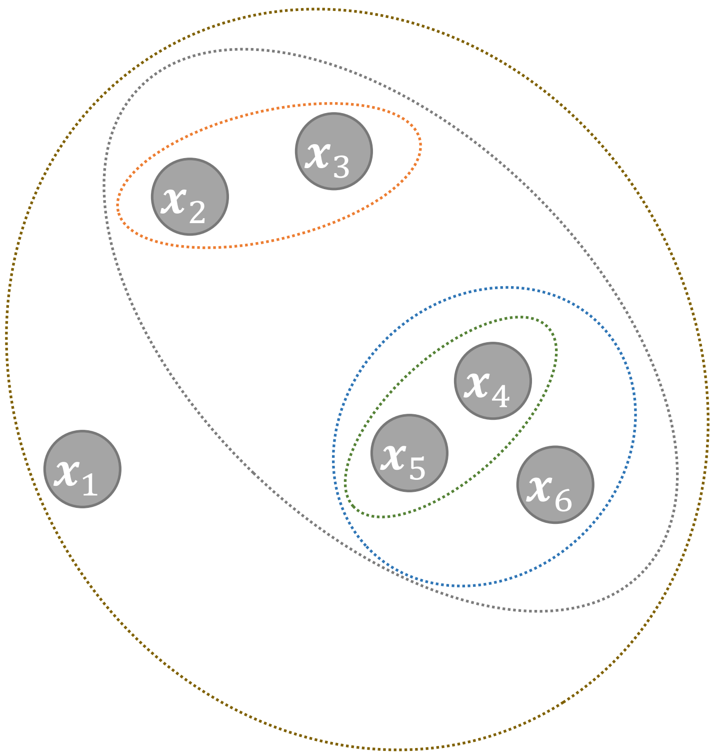
Difference to classification:
Groups not given!

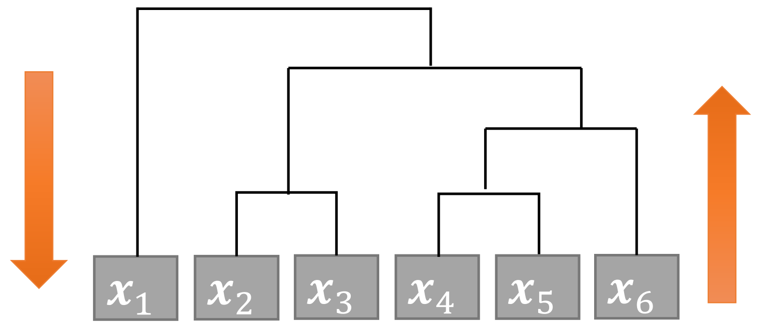
Task



Variations of the Task

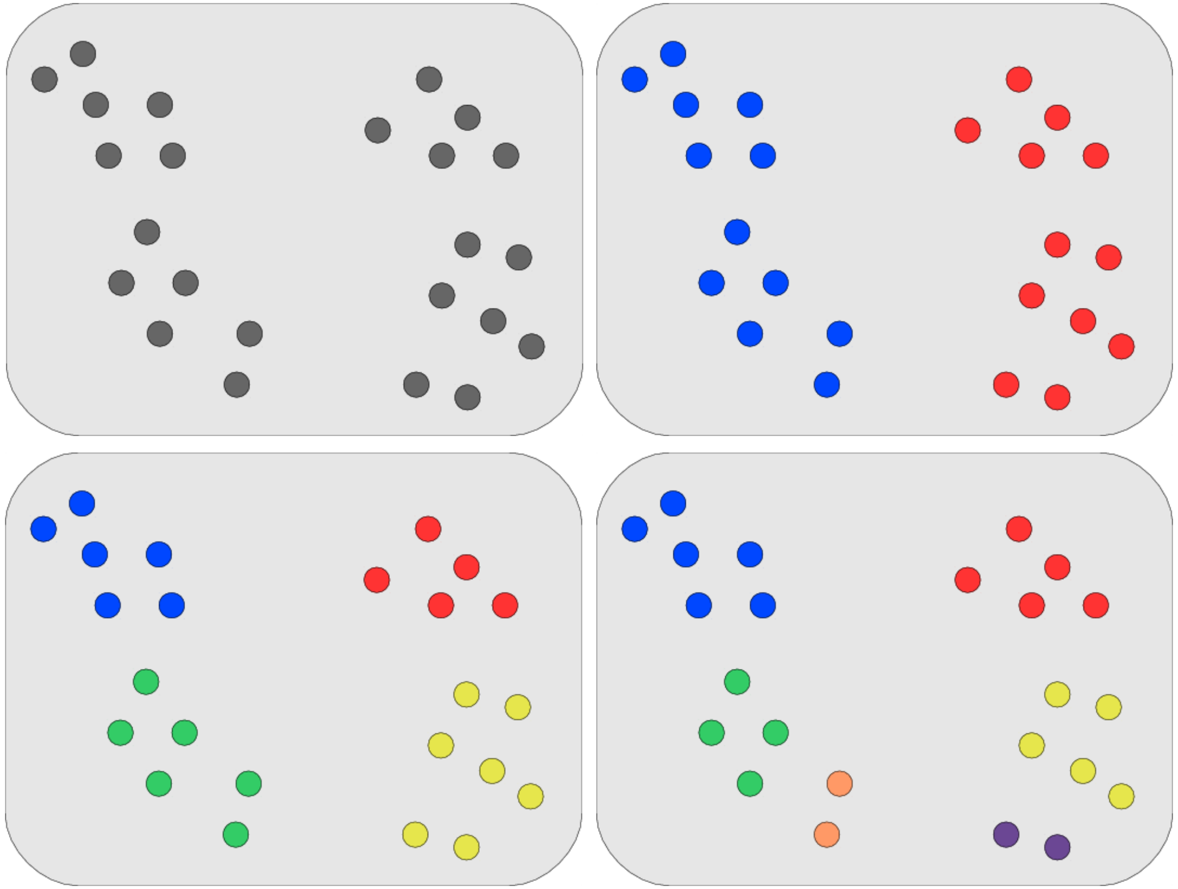
- Cluster types:
  - Flat vs. Hierarchical
  - Exclusive vs. Multiple clusters
    - $\gamma: D \to \wp(\Psi)$
- Function γ
  - Hard vs. Soft assignments
    - $\gamma: D \to \Psi \times \mathbb{R}$
  - Based on shape, density, estimates of distribution mixture





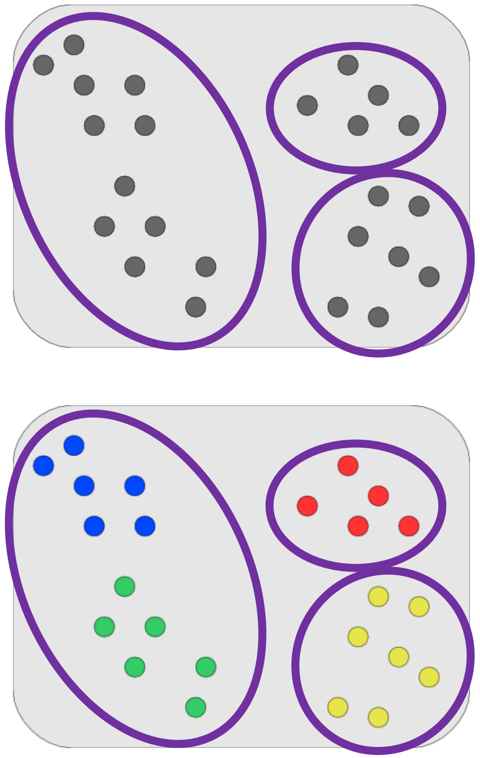
Cardinality

- Number of clusters
  - Provided (externally) or to be defined (given explicit hyperparameter) or found over the data (given other hyperparameters, e.g. density or density distribution).



Evaluation

- Intrinsic
  - Evaluate the quality of clusters directly.
    - E.g. compactness, separation of groups, etc.
- Extrinsic
  - Employ external knowledge
    - Ground truth from classification data
    - Assuming categories to be optimal clusters
  - Compare found with pre-defined clusters
    - Difficulty of finding a matching
- Indirect
  - User testing (satisfaction, task performance)
  - Application specific metrics

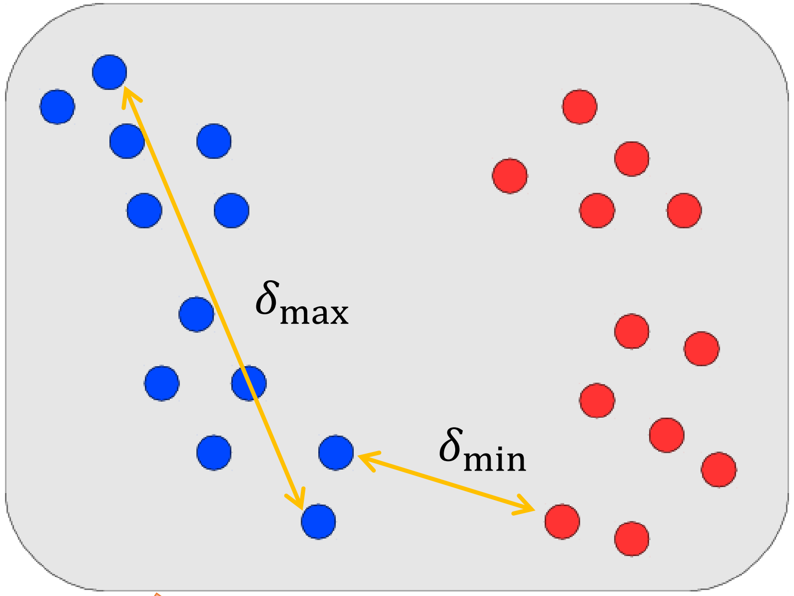


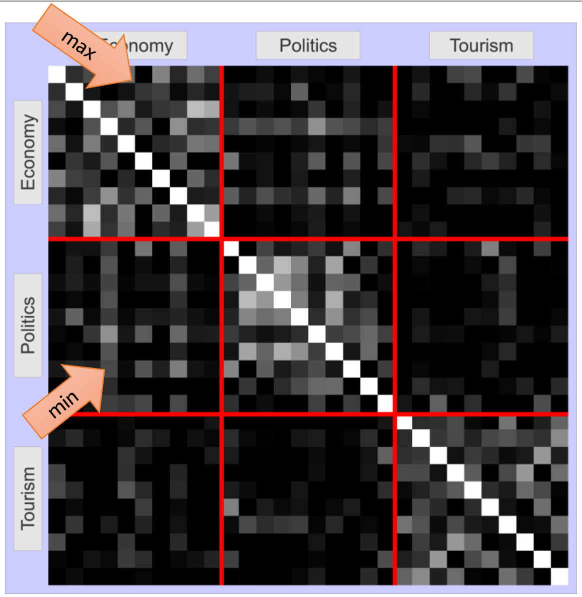
- Dunn Index
  - Notion of cluster separation

$$I_{\text{Dunn}}(\Psi) = \frac{\delta_{\min}}{\delta_{\max}}$$

    - $\delta_{\min}$ smallest inter-cluster distance
    - $\delta_{\max}$ largest intra-cluster distance
- Requires pair-wise distances
  - Distance matrix
  - Graphical representation
    - Minimal distance: white
    - Maximal distance: black
- Applicable also to ground truth
  - Notion of difficulty of cluster problem
  - Example

$$I_{\text{Dunn}}(\{c_E, c_P, c_T\}) = \frac{0.577}{1.414} = 0.435$$





- Silhouette coefficient $s(i)$ for object $\boldsymbol{x}_i$
  - Average distance $a(i)$ to all other objects in the same cluster ψ

$$a(i) = \sum_{\boldsymbol{x} \in \psi} \frac{1}{|\psi|} \delta(\boldsymbol{x}_i, \boldsymbol{x})$$

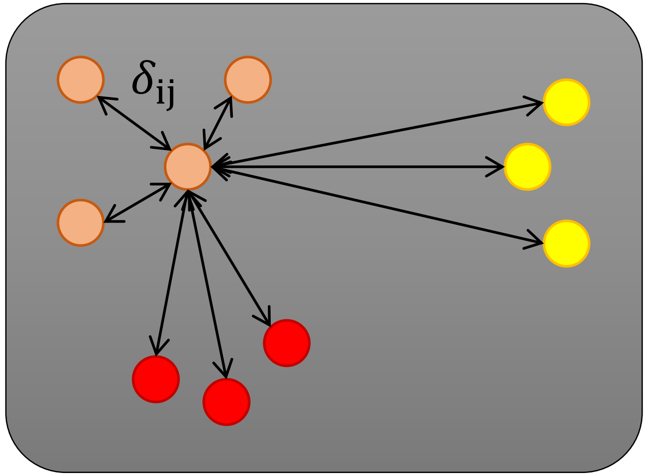
  - Average distance to any other cluster $\psi', \psi' \neq \psi$:

$$d(i, \psi') = \sum_{\boldsymbol{x} \in \psi'} \frac{1}{|\psi'|} \delta(\boldsymbol{x}_i, \boldsymbol{x})$$

  - Average distance $b(i)$ to the closest cluster $\psi', \psi' \neq \psi$:

$$b(i) = \min_{\psi' \epsilon \Psi} d(i, \psi')$$

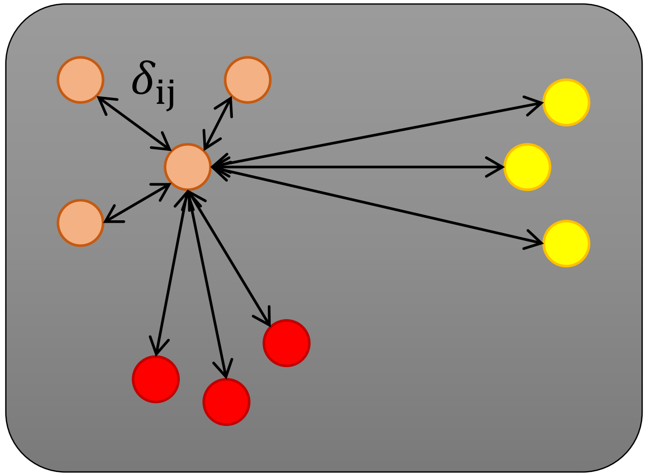
  - Silhouette coefficient: $s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$



Intrinsic metrics

- Silhouette coefficient
  - Values: $-1 \leq s(i) \leq 1$
  - Value close to 1:
    - $a(i)$ much smaller than $b(i)$
    - Distances within cluster very small in comparison to distances with other clusters
  - Value close to 0:
    - $a(i) \approx b(i)$
    - Same internal as external distance
  - Value close to -1:
    - $b(i)$ much smaller than $a(i)$
    - Other instances are (on average) closer than the instances in same cluster
- Aggregation: Average silhouette coefficient

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$



Extrinsic metrics

- External knowledge (ground truth of categories)

$$\Omega = \{\omega_1, \cdots, \omega_{k'}\}$$

- Compare the clusters Ψ and categories Ω
- Determine:
  - $n_j^{(u)}$: number of objects from ω_u being clustered into ψ_j
- Purity
  - Ratio of strongest represented category

$$\mathrm{Purity}(\psi_j) = \frac{1}{|\psi_j|} \cdot \max_{u=1,\cdots,k'} n_j^{(u)}$$

  - Aggregate over all clusters

$$\mathrm{Purity}(\Psi) = \sum_{j=1}^{k} \frac{|\psi_j|}{N} \cdot \mathrm{Purity}(\psi_j)$$

Purity of 1 can always be achieved!

Extrinsic metrics

- Mutual Information:
  - Mutual agreement between clusters and categories

$$\mathrm{MI}(\Psi) = \frac{1}{N}\sum_{j=1}^{k}\sum_{u=1}^{k\prime} n_j^{(u)} \cdot \log \frac{n_j^{(u)} \cdot N}{\sum_{m=1}^{k\prime} n_j^{(m)} \cdot \sum_{t=1}^{k} n_t^{(u)}}$$

  - Log base: 2 or $k \cdot k'$
- Rand Index:
  - Consider document pairs on categories and clusters
    - Agreements: same-same (ss), different-different (dd)
    - Disagreements: same-different (sd), different-same (ds)
  - Agreement-ratio:

$$I_{\mathrm{Rand}}(\Psi) = \frac{\mathrm{ss} + \mathrm{dd}}{\mathrm{ss} + \mathrm{dd} + \mathrm{sd} + \mathrm{ds}}$$

Example

1. $\psi_1 = \{\boldsymbol{x_1, x_2, x_3, x_4, x, x_6, x_7, x_8, x_9, x_{10}}, \boldsymbol{x_{11}, x_{20}}\}$
2. $\psi_2 = \{\boldsymbol{o_{12}, o_{13}, o_{14}, o_{15}, o_{16}, o_{17}, o_{18}, o_{19}}\}$
3. $\psi_3 = \{\boldsymbol{x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{27}, x_{28}, x_{30}}\}$
4. $\psi_4 = \{\boldsymbol{x_{26}, x_{29}}\}$

- Purity:

$$\text{Purity}(\psi_1) = \frac{10}{12} = 0.83$$

$$\text{Purity}(\Psi) = \frac{12}{30} \cdot 0.83 + \frac{8}{30} \cdot 1.0 + \frac{8}{30} \cdot 1.0 + \frac{2}{30} \cdot 1.0 = 0.93$$

- Mutual Information
  - Log base $k \cdot k'$
  - Several values are 0

$$\text{MI}(\Psi) = \frac{1}{30} \cdot \left(10 \cdot \log \frac{10 \cdot 30}{12 \cdot 10} + 2 \cdot \log \frac{2 \cdot 30}{12 \cdot 10} + 8 \cdot \log \frac{8 \cdot 30}{8 \cdot 10} + 8 \cdot \log \frac{8 \cdot 30}{8 \cdot 10} + 2 \cdot \log \frac{2 \cdot 30}{2 \cdot 10}\right) = 0.370$$

Example

1. $\psi_1 = \{\boxed{x_1, x_2, x_3, x_4, x, x_6, x_7, x_8, x_9, x_{10}}, \boxed{x_{11}, x_{20}}\}$
2. $\psi_2 = \{\boxed{o_{12}, o_{13}, o_{14}, o_{15}, o_{16}, o_{17}, o_{18}, o_{19}}\}$
3. $\psi_3 = \{\boxed{x_{21}, x_{22}, x_{23}, x_{24}, x_{25}, x_{27}, x_{28}, x_{30}}\}$
4. $\psi_4 = \{\boxed{x_{26}, x_{29}}\}$

- Agreements:
  - $ss = 103 = \binom{10}{2} + 1 + \binom{8}{2} + \binom{8}{2} + 1$
  - $dd = 280 = 10 \cdot 18 + 2 \cdot 10 + 8 \cdot 10$
- Disagreements
  - $sd = 32 = 2 \cdot 8 + 2 \cdot 8$
  - $ds = 20 = 2 \cdot 10$
- Rand Index:

$$I_{\text{Rand}}(\Psi) = \frac{383}{435} = 0.88$$

K-Means

K-Means

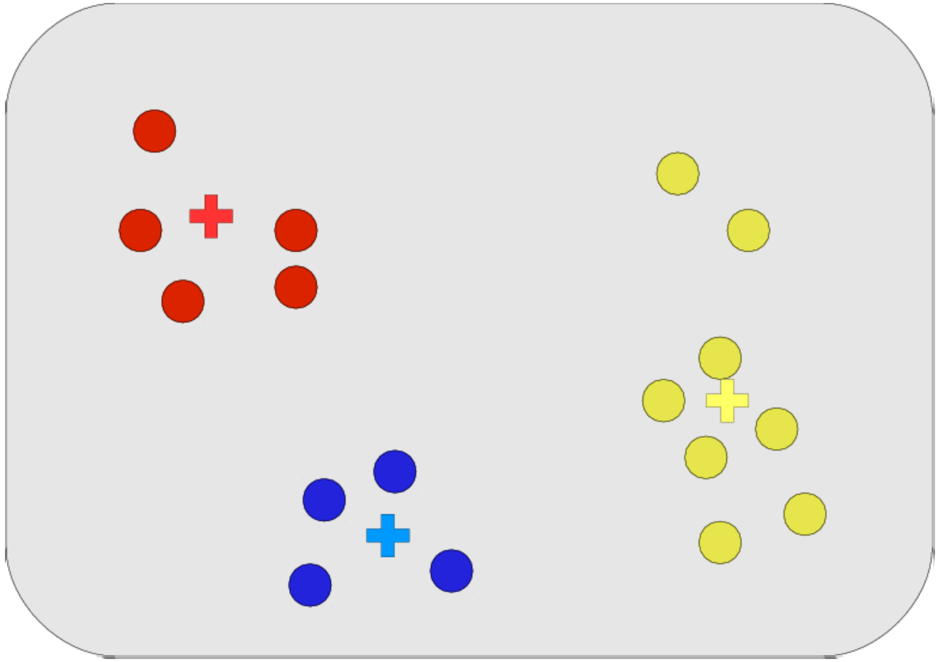
- General clustering algorithm

- Characteristics:
  - Flat clusters
  - No overlaps
  - Good runtime
  - Simple to implement

- Parameters
  - k : number of clusters
  - Initial random seed

Algorithm

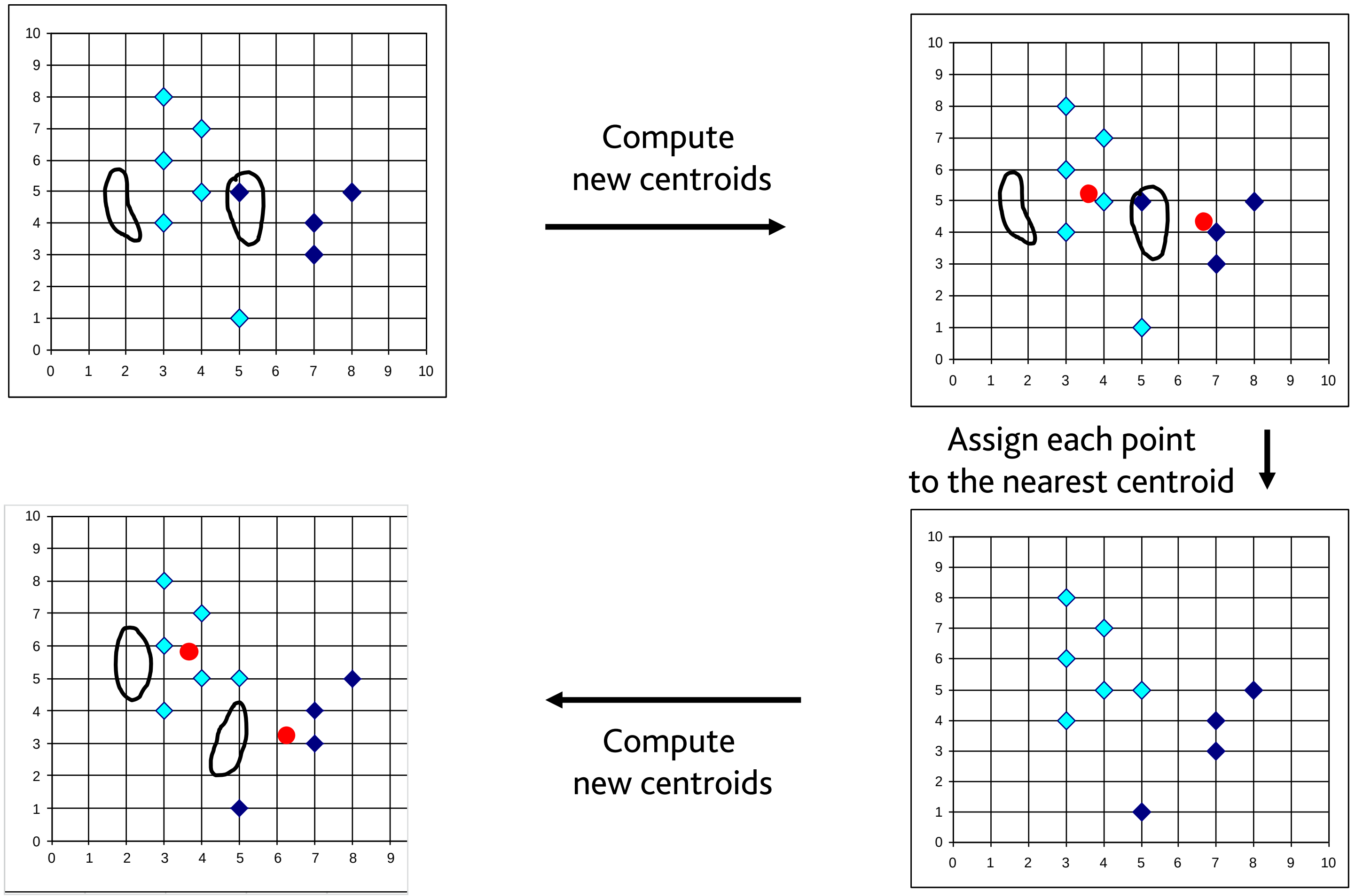
- Random seed Ψ for k clusters (e.g. single objects)
- Determine centroids Z for clusters
- While centroids not stable
  - For all objects $\boldsymbol{x}_i$
    - Reassign $\boldsymbol{x}_i$ to cluster ψ_j with minimal $\delta\left(\overrightarrow{\boldsymbol{x}_i}, \vec{z}_j\right)$
  - For all clusters ψ_j
    - Re-compute centroid $\vec{z}_j$



K-means Algorithm

```
K-means(Set_of_points D, Integer k)
   Create an "initial" partitioning of D in k cluster;
   Compute a set Z'={Z'_1, ..., Z'_k} of centroids for the k Cluster;
   Z = {};
   repeat until Z = Z'
      Z = Z';
      Generate k clusters by assigning each data point to the nearest centroid
        Z_j;
      Compute the set Z'={Z'_1, ..., Z'_k} of centroids for the new clusters;
   return Z';
```

K-Means: example



Advantages and disadvantages of k-means

- Advantages:
  - Efficiency:
    time complexity: $O(n)$ for each iteration,
    Number of iterations is usually very small (~ 5 - 10).
  - Simple implementation
  - Easy, good interpretability
  - *K*-means is the most popular clustering algorithm!

- Disadvantages:
  - Susceptible to noise and outliers since all objects influence the computation of centroids
  - Cluster have always convex form
  - The number of clusters k is often difficult to determine
  - Strong dependency on initial partition (runtime + result!)

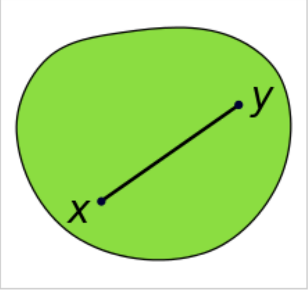


Illustration of a convex set which looks somewhat like a deformed circle. The (black) line segment joining points x and y lies completely within the (green) set. Since this is true for any points x and y within the set that we might choose, the set is convex.

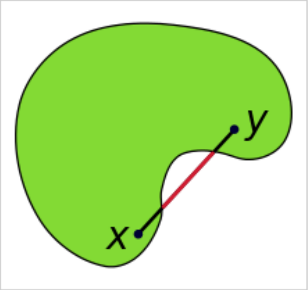


Illustration of a non-convex set. Since the red part of the (black and red) line-segment joining the points x and y lies *outside* of the (green) set, the set is non-convex.

Variations

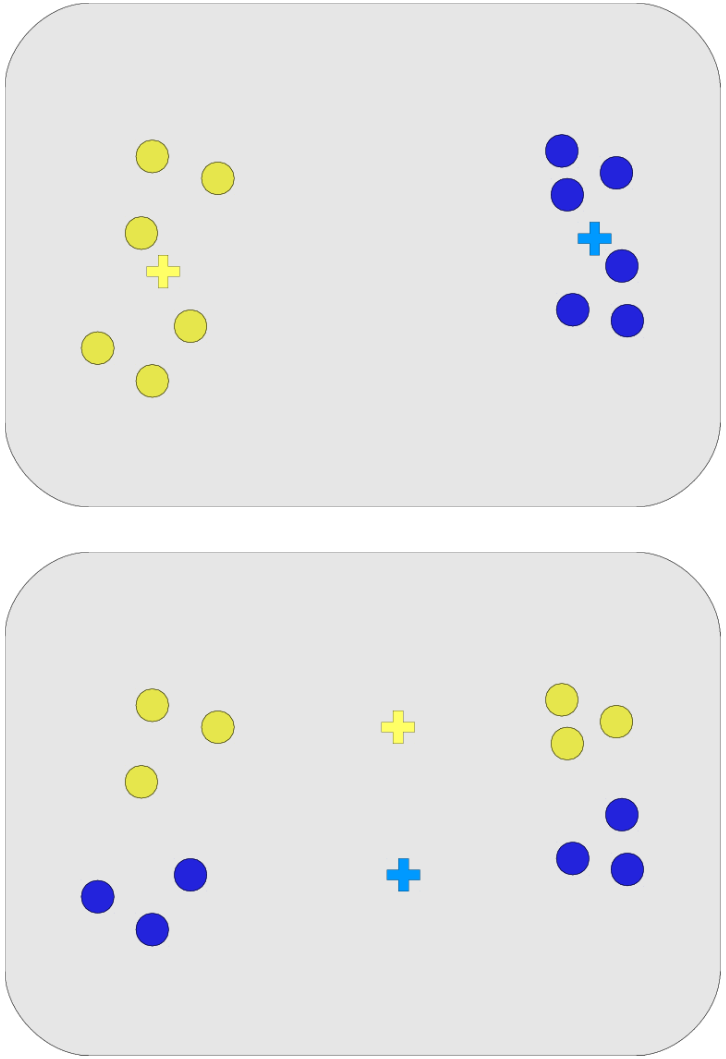
- Random seed
  - Assign all objects
  - Compute actual centroids
- Stop criterion
  - No change of clusters (equivalent)
  - Small changes of the centroids
  - Fixed number of iterations
- Centroids
  - Non-standard metrics (e.g. string similarity)
    - Mean centroid cannot be computed
  - K-Median
    - Use most central objects

Initial configuration

- Choice of initial seed can cause different outcomes!
- Solution
  - Repeat with different seeds
  - Evaluate quality
    - Dunn index
    - Residual Sum of Squares

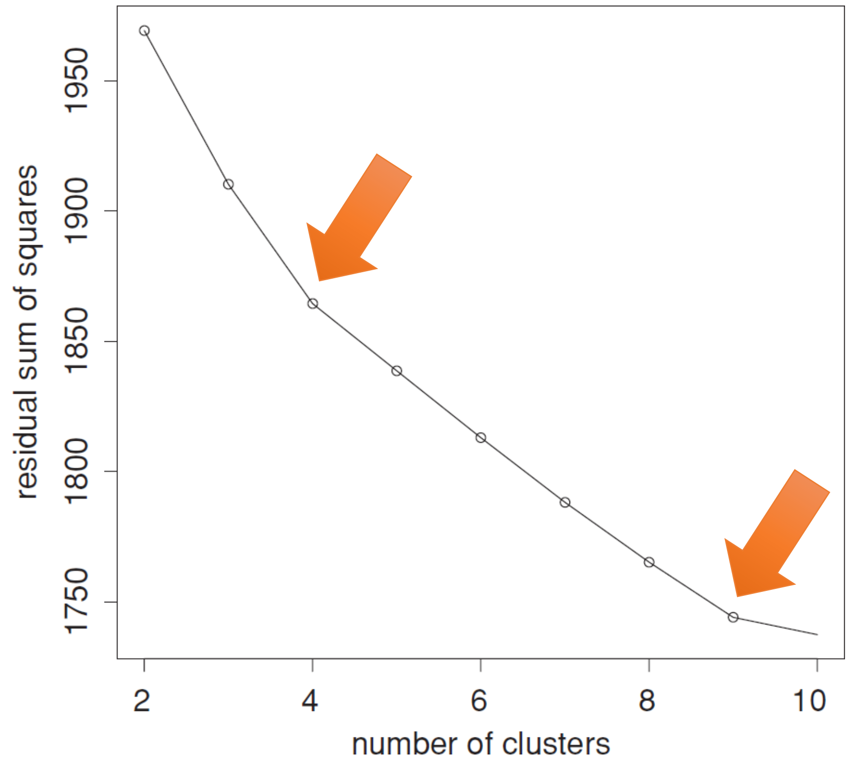
$$RSS(\Psi) = \sum_{j=1}^{k} \sum_{d_i \in \psi_j} \delta\left(\vec{d}_i, \vec{z}_j\right)^2$$

- Choose best performing setting



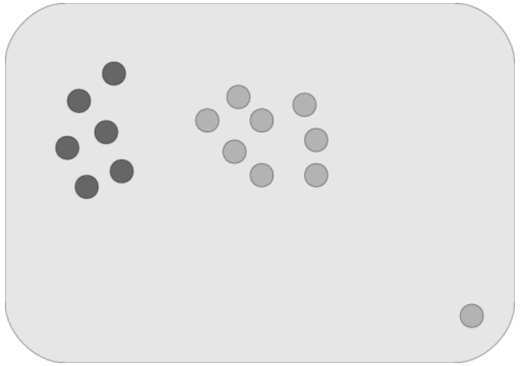
Choice of K

- Important parameter!
- Knowledge about the data
  - Expert insights
- Development of RSS
  - Monotonous decline
  - Typically two points where decline slows down

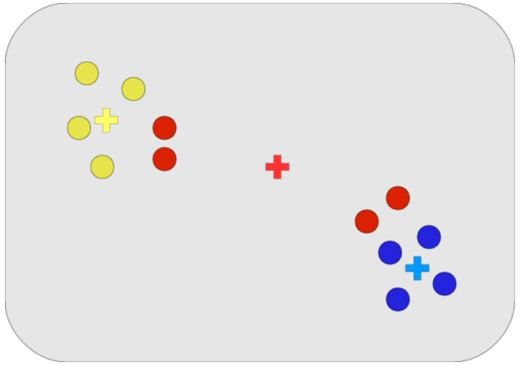


Problematic configurations

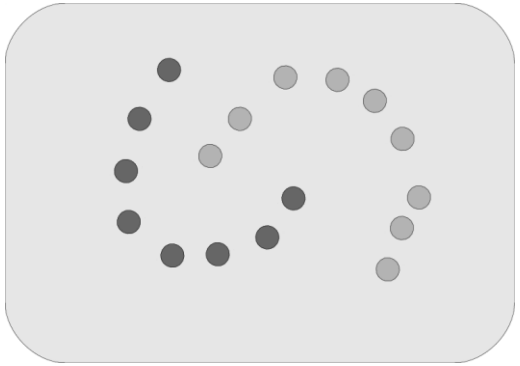
- Outliers
  - Cause singleton clusters
  - Solution:
    - Remove and treat separately

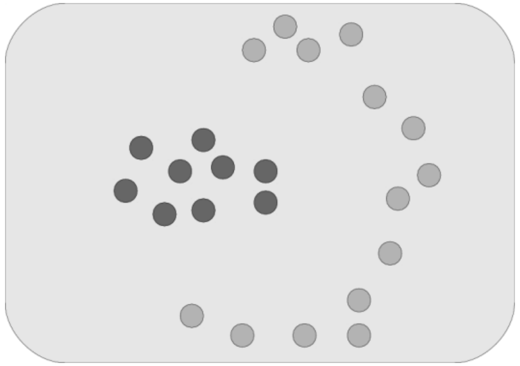


- Empty clusters
  - Unlucky position of centroids
  - Solution:
    - Split large cluster

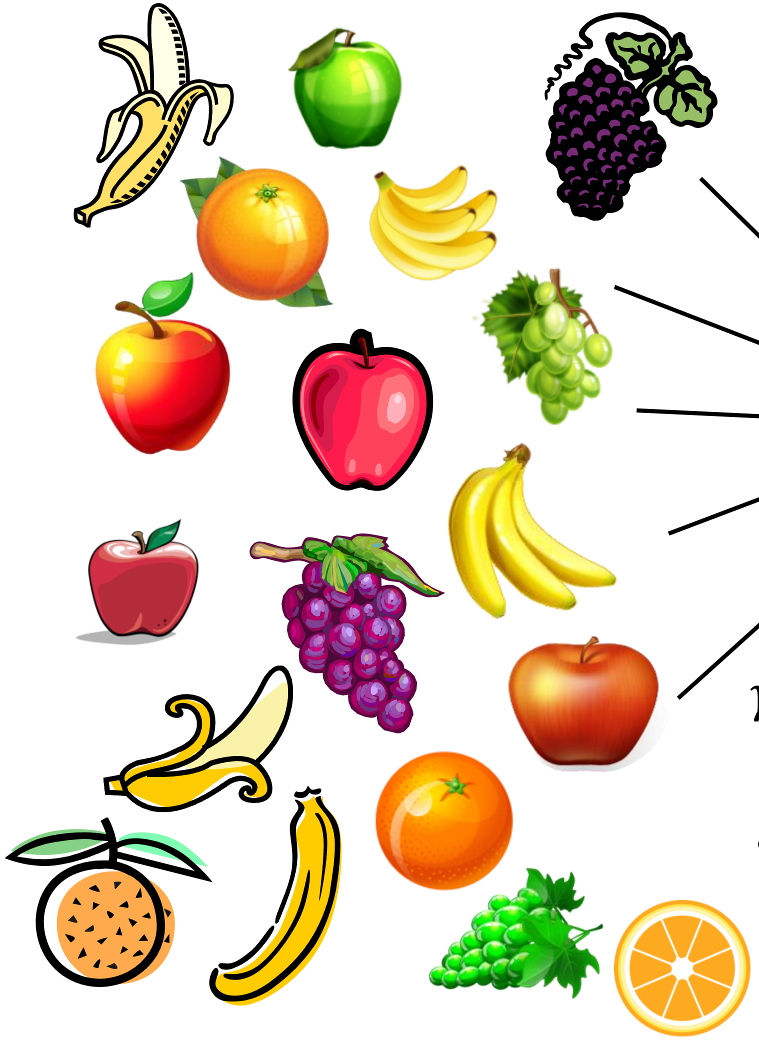


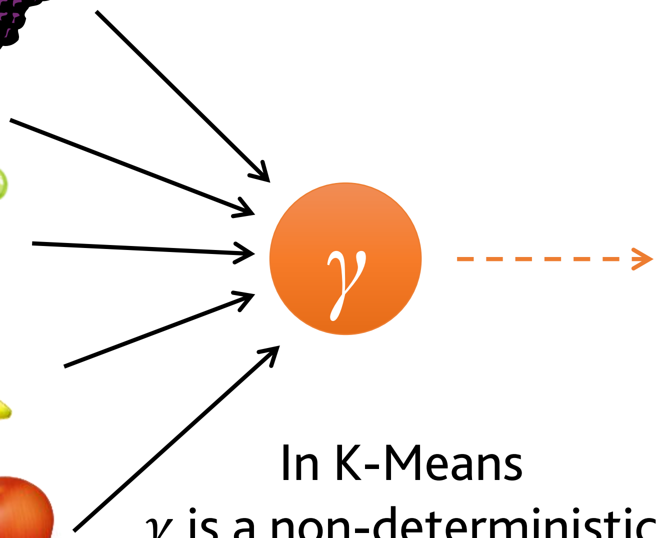
- Non-spheric shapes
  - Cannot be handled!

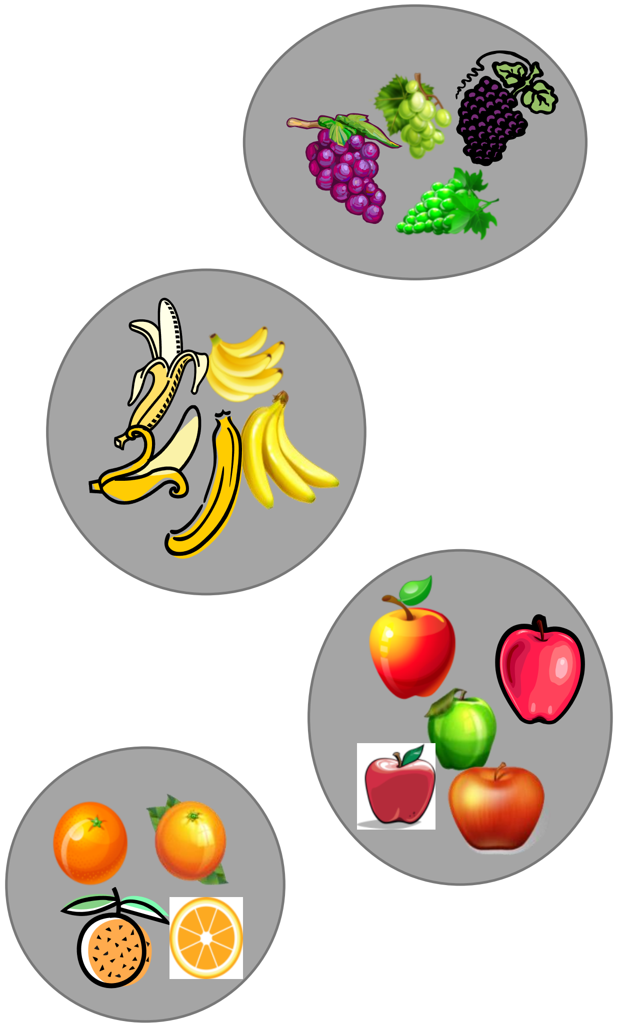




K-Means-Clustering







In K-Means
γ is a non-deterministic function defined by greedy heuristic search that optimizes the sum-of-squared distances

Probabilistic Models

Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

- Example: Throwing a coin: 10 x head, 2 x tail

Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

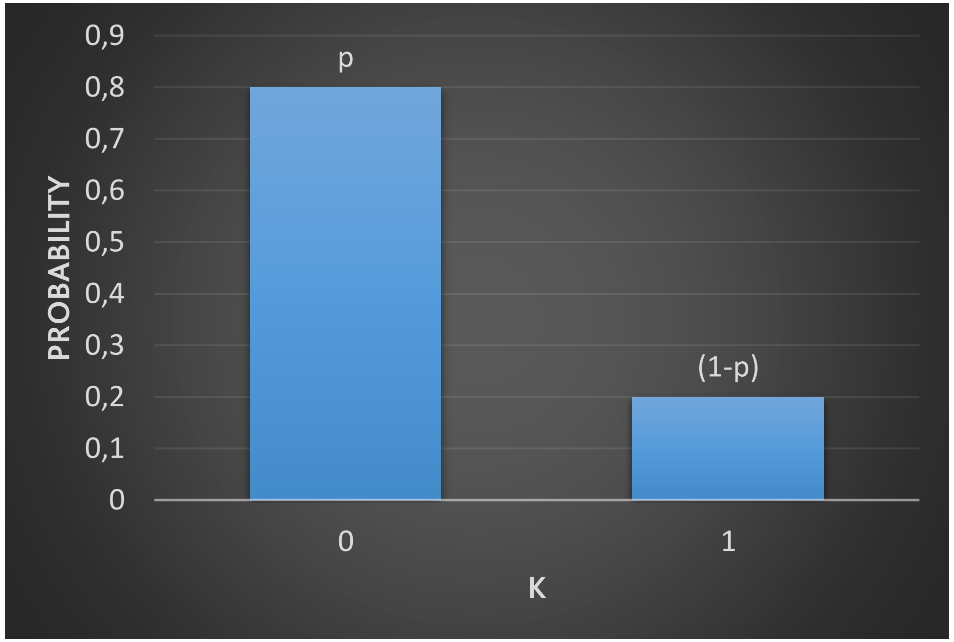
- Example: Throwing a coin: 8 x head, 2 x tail

Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

- Example: Throwing a coin: 8 x head, 2 x tail

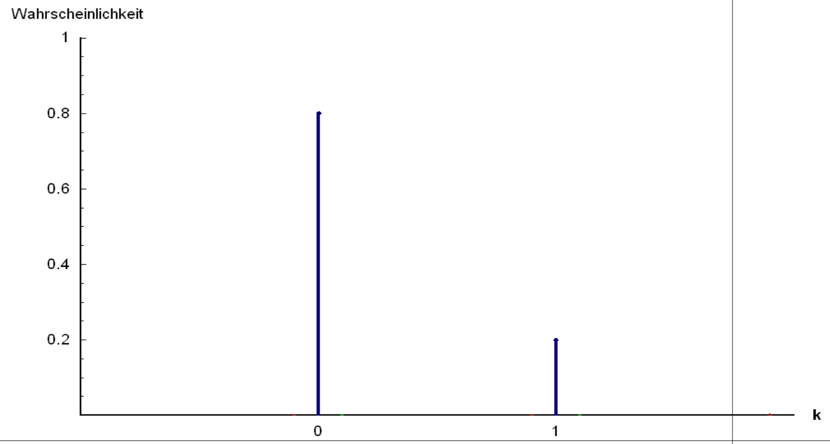
- θ: Bernoulli (p)



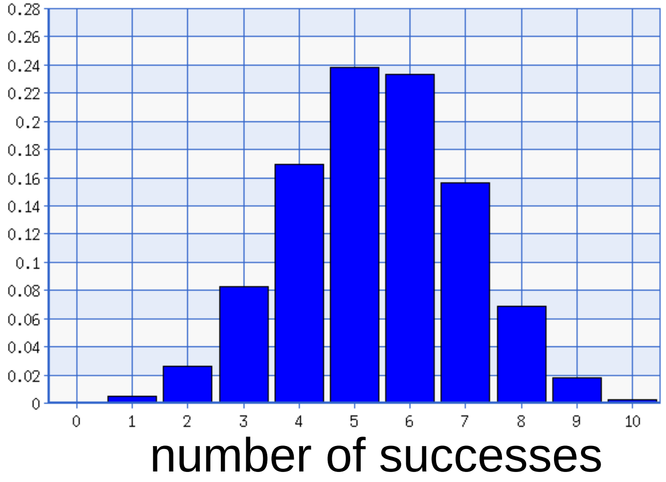
Overview: Probability distributions

	n=1	n >= 1
k = 2	Bernoulli = Binomial for n=1	Binomial
k > 2	Categorical	Multinomial

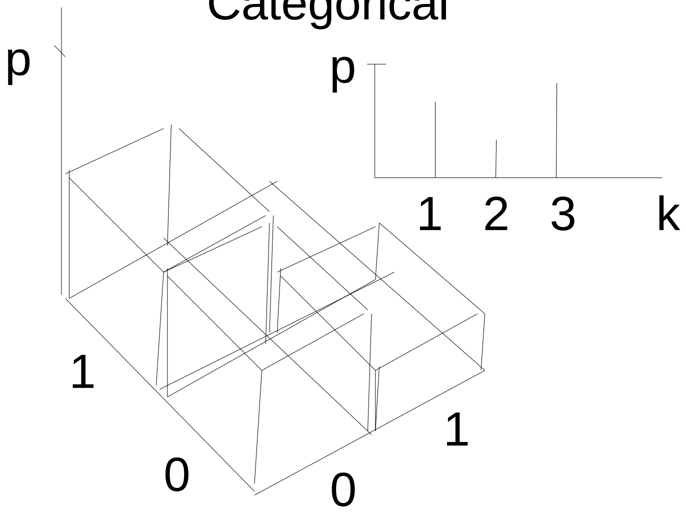
Bernoulli = Binomial for n=1



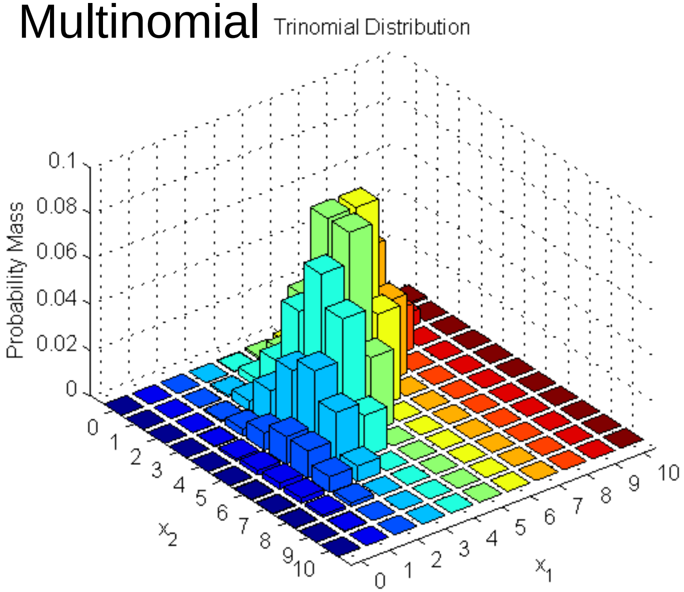
Binomial



Categorical



Multinomial



Overview: Probability distributions

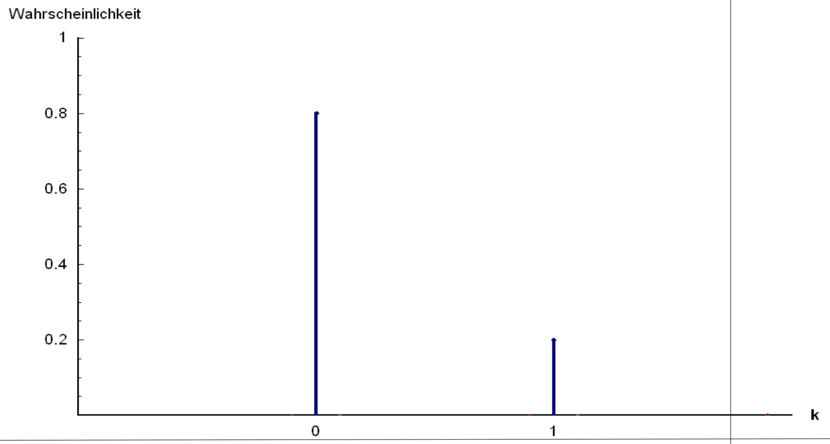
n=1

n >= 1

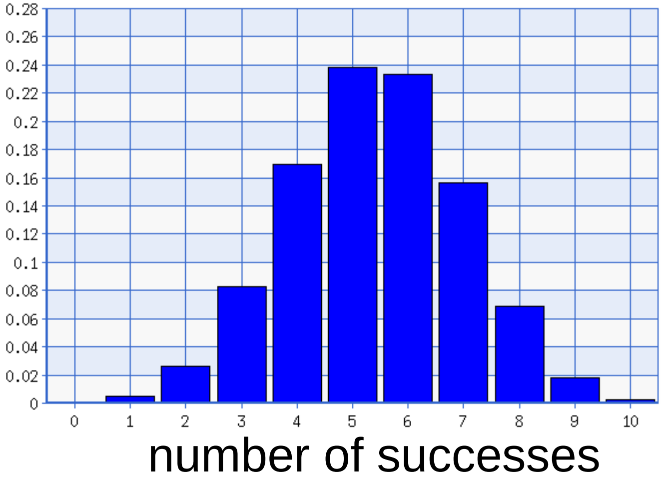
n $\rightarrow \infty$

k = 2

Bernoulli = Binomial for n=1



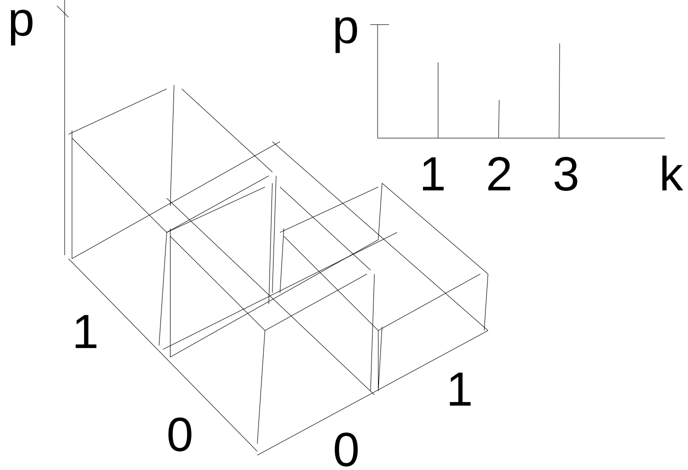
Binomial



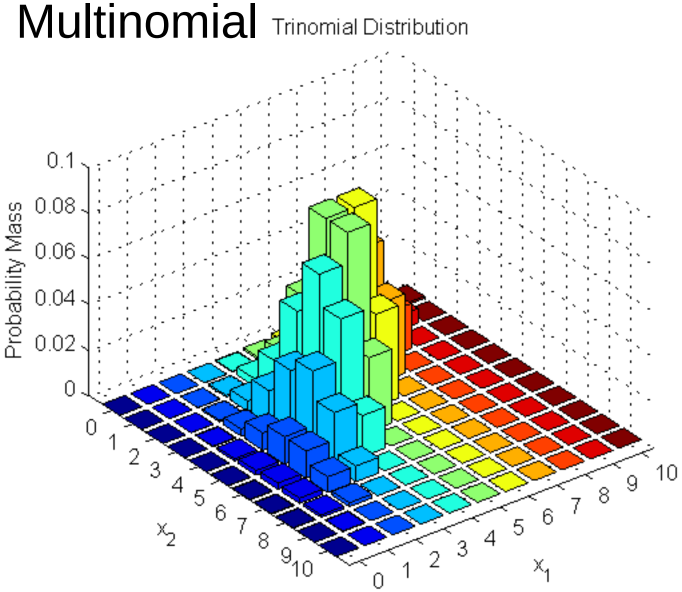
Gaussian

k > 2

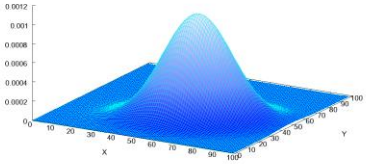
Categorical



Multinomial



Multivariate Gaussian

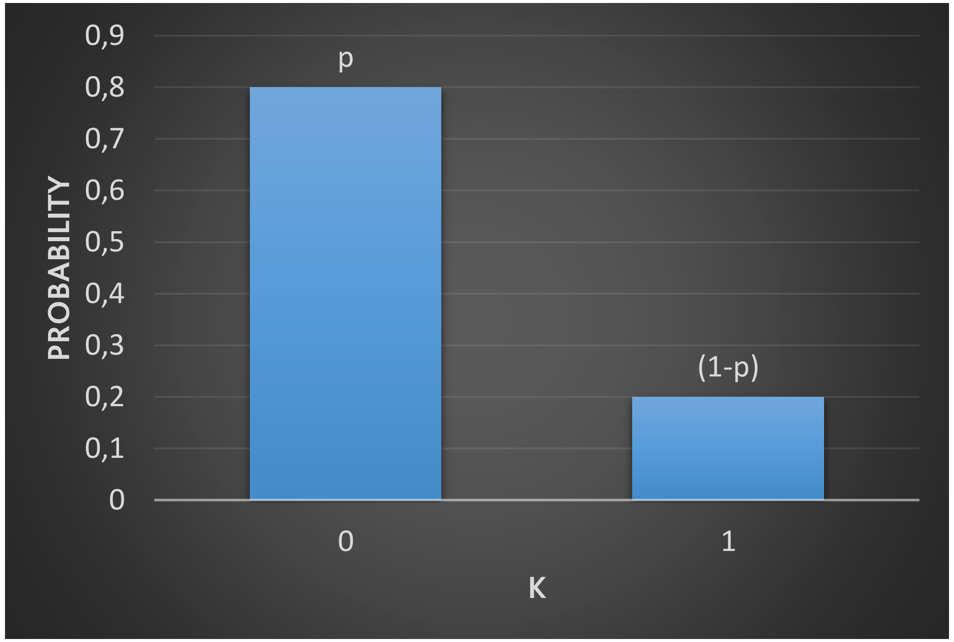


Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

- Example: Throwing a coin: 8 x head, 2 x tail

- θ: Bernoulli (p)



How to estimate p?

Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

- Example: Throwing a coin: 8 x head, 2 x tail

- θ: Bernoulli (p)

- Likelihood: $L = P(D|\theta) = \prod_{d_i \in D} P(d_i|\theta)$

Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

- Example: Throwing a coin: 8 x head, 2 x tail

- θ: Bernoulli (p)

- Likelihood: $L = P(D|\theta) = \prod_{d_i \in D} P(d_i|\theta)$

- Log-Likelihood: $\sum_{d_i \in D} \log(P(d_i|\theta))$

Probabilistic models

- Observations D (our Data)
- Hidden (latent) parameters θ

- Example: Throwing a coin: 8 x head, 2 x tail

- θ: Bernoulli (p)

- Likelihood: $L = P(D|\theta) = \prod_{d_i \in D} P(d_i|\theta)$

- Log-Likelihood: $\sum_{d_i \in D} \log(P(d_i|p)) \quad |d_i \in \{H, T\}$

Probabilistic models

- θ: Bernoulli (p)

- Likelihood: $L = P(D|\theta) = \prod_{d_i \in D} P(d_i|\theta)$

- Log-Likelihood: $\sum_{d_i \in D} \log(P(d_i|p)) \quad |d_i \in \{H, T\}$

$$= n^T \cdot \log\big(P(T|p)\big) + n^H \cdot \log\big(P(H|p)\big)$$

$$= n^T \cdot \log(p) + n^H \cdot \log(1 - p)$$

Probabilistic models

$$\log L = n^T \cdot \log(p) + n^H \cdot \log(1-p)$$

- Maximization:

$$\frac{\partial \log L}{\partial p} = \frac{n^T}{p} - \frac{n^H}{1-p} = 0$$

$$\Leftrightarrow p = \frac{n^T}{n^T + n^H} = \frac{2}{2+8} = 0.2$$

Expectation Maximization

Expectation Maximization (EM)

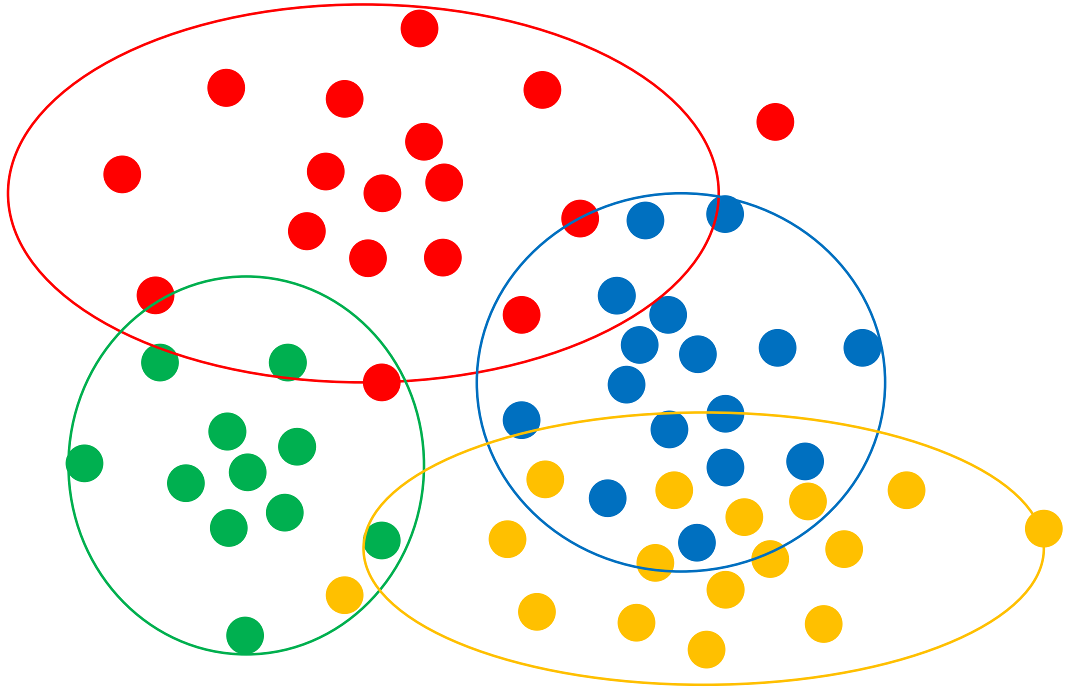
- General clustering algorithm

- Characteristics:
  - Probabilistic approach
  - Soft assignments to clusters
  - Generalization of K-Means

- Parameters
  - K : number of clusters
  - Initial random seed
  - Model for the distribution

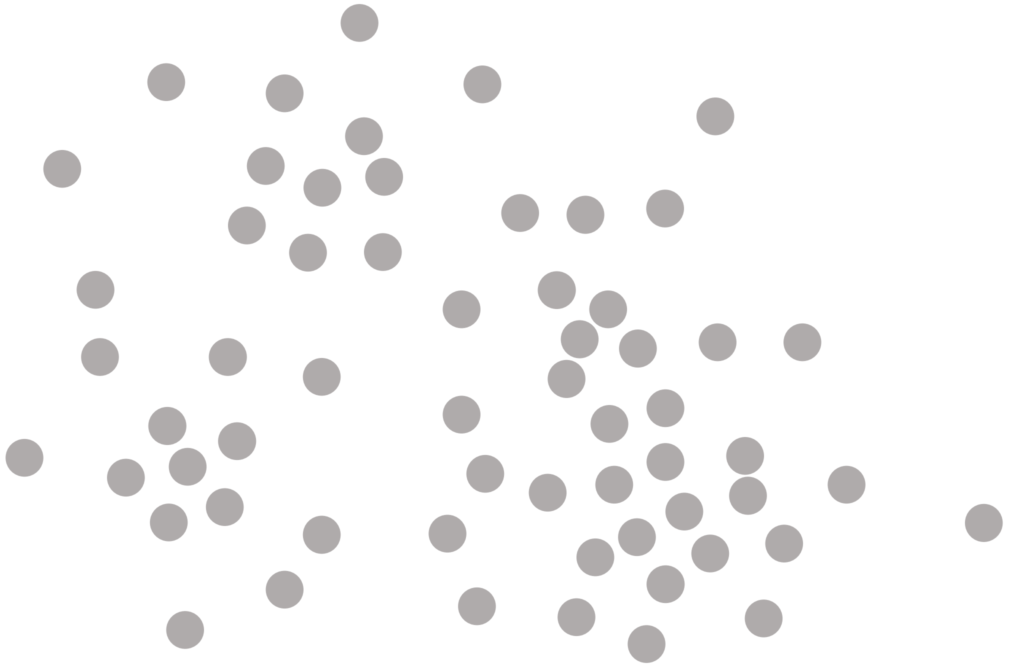
Idea

- Data can be explained by a mixture of parametrized probability distributions – one per cluster.



Idea

- Data can be explained by a mixture of parametrized probability distributions – one per cluster.
- Problem: the true distributions are unknown, all we see is the data:



Model of the distribution

- Provide a model how the data is distributed, e.g.
  - Density of a normal distribution (1 dimensional)
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$
    - Where μ is the mean and σ the standard deviation
  - Density of a normal distribution (m-dimensional)
$$f(\boldsymbol{x}) = \frac{1}{\sigma\sqrt{(2\pi)^m |S|}} e^{-\left(\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T S(\boldsymbol{x}-\boldsymbol{\mu})\right)}$$
    - Where μ is an m-dimensional vector, S an m×m covariance matrix and $|S|$ the determinant
  - Other distribution P with parameters ϑ

Model estimation

- Estimate model parameters from data
- Maximum likelihood estimation:
  - 1 dimensional case:

$$\mu = \frac{1}{n}\sum_{x\in D} x_i \qquad \sigma = \sqrt{\frac{1}{n-1}\sum_{x\in D}(x_i-\mu)^2}$$

  - m-dimensional case:

$$\vec{\mu} = \frac{1}{n}\sum_{x\in D} \vec{\boldsymbol{x}_i} \qquad S = \frac{1}{n-1}\sum_{x\in D}(\vec{\boldsymbol{x}_i}-\vec{\mu})(\vec{\boldsymbol{x}_i}-\vec{\mu})^T$$

  - Maximize log likelihood function:

$$\log(L(\vartheta|D)) = \log\Big(\prod_{x\in D} P(\boldsymbol{x}|\vartheta)\Big) = \sum_{x\in D}\log(P(\boldsymbol{x}|\vartheta))$$

But: we don't know which objects belongs to which cluster ...

Latent variables: Cluster assignment

- For each object $\boldsymbol{x}_i \in D$, we have a latent variables z_{ij} modelling to which cluster ω_j it belong
- Two steps:
  - Expectation step
    - Calculate the expected values for z_{ij} given the current model parameters ϑ
    - So: how probable is it that object $\boldsymbol{x}_i$ belongs to cluster ω_j under the current model hypothesis ϑ
  - Maximization step
    - Calculate the model parameters ϑ given the current estimates for the latent variables z_{ij}
    - So: how do the model parameters look like under the assumption that the assigment to clusters is correct.
- Iterate until convergence (little change)

Initialization

- Problem:
  - Expectation step needs model parameters to estimate latent variables
  - Maximization step needs latent variables to estimate model parameters

- Different options:
  - Hand selected initial model parameters
  - Random initialization
  - Choose random individuals
  - Perform k-means cluster to find initial clusters

Example

- Cluster people by their height
  - Two classes
  - Assume initial model:
    - $\mu_1 = 110$ $\sigma_1 = 20$
    - $\mu_2 = 160$ $\sigma_2 = 20$

- Expectation:
  - $f_{\mu_1,\sigma_1}(124) = 0.01561$
  - $f_{\mu_2,\sigma_2}(124) = 0.00395$
  - Weights:
    - $z_{1,1} = 0.7982$
    - $z_{1,2} = 0.2018$
  - …

Gender	height
F	124
F	115
F	121
F	139
F	98
F	135
F	131
M	170
M	166
M	155
M	167
M	158
M	175
M	143
M	163
M	160
M	145
M	176

Example

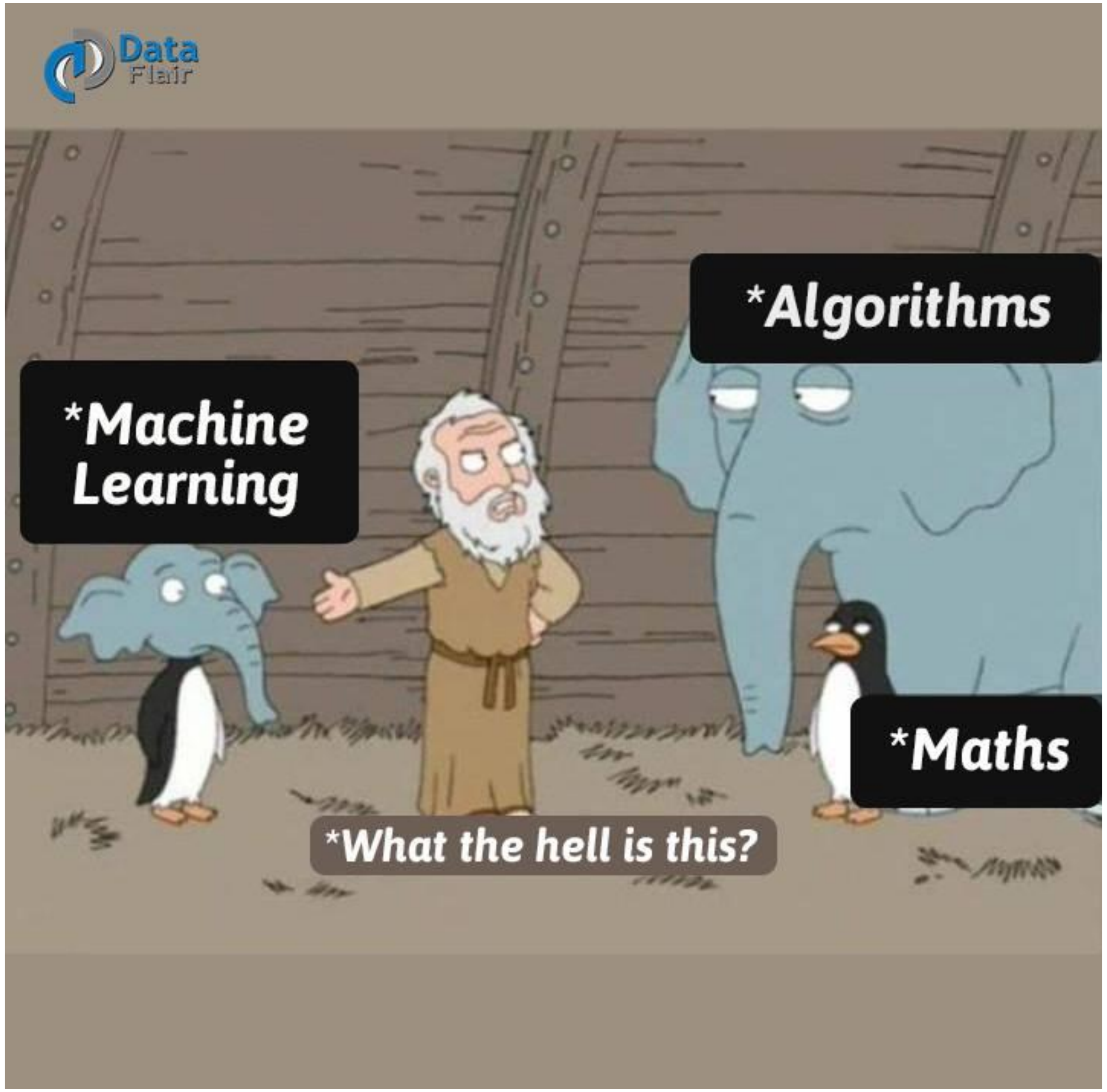
- Given all weights:
- New model parameters

$$\mu_j = \frac{\sum_{x \in D} z_{i,j}\, \boldsymbol{x}_i}{\sum_{x \in D} z_{i,j}}$$

$$\sigma_j = \sqrt{\frac{\sum_{x \in D} z_{i,j} (\boldsymbol{x}_i - \mu)^2}{\sum_{x \in D} z_{i,j}}}$$

- Values:
  - $\mu_1 = 123.72$ $\sigma_1 = 15.98$
  - $\mu_2 = 157.72$ $\sigma_2 = 14.62$

Gender	height
F	124
F	115
F	121
F	139
F	98
F	135
F	131
M	170
M	166
M	155
M	167
M	158
M	175
M	143
M	163
M	160
M	145
M	176


Data Flair
*Algorithms
*Machine Learning
*Maths
*What the hell is this?

Summary

Summary

- Clustering
  - K-Means
  - Expectation-Maximization algorithm

Thank you!

Zeyd Boukhers

E-mail: Boukhers@uni-koblenz.de
Phone: +49 (0) 261 287-2765
Web: Zeyd.Boukhers.com

University of Koblenz-Landau
Universitätsstr. 1
56070 Koblenz