



Big Data Analytics

– Week 2: k-Means Clustering –

Knowledge Discovery and Data Mining
Prof. Dr. Emmanuel Müller

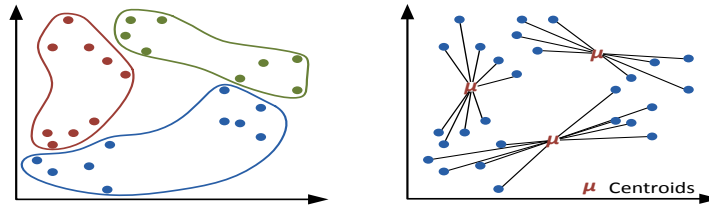
K-Means Clustering: Basic Idea

S.P. Lloyd: Least squares quantization in PCM. In IEEE Information Theory, 1982 (technical report, Bell Labs, 1957)
 J. MacQueen: *Some methods for classification and analysis of multivariate observation*,
 In Proc. of the 5th Berkeley Symp. on Math. Statist. and Prob., 1967.

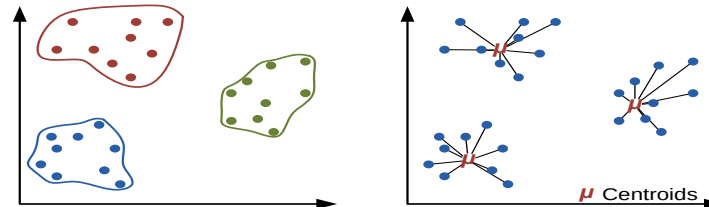
Idea of K-means: find a clustering such that **the within-cluster variation** of each cluster is small and use the **centroid** of a cluster as representative.

Objective: For a given k , form k groups so that the sum of the (squared) distances between the mean of the groups and their elements is minimal.

Poor Clustering
(large sum of distances)



Optimal Clustering
(minimal sum of distances)



K-Means Clustering: Basic Notions

Objects $p = (p_1, \dots, p_d)$ are points in a d -dimensional vector space
(the mean μ_S of a set of points S must be defined: $\mu_S = \frac{1}{|S|} \sum_{p \in S} p$)

Measure for the compactness of a cluster C_j (sum of squared errors):

$$SSE(C_j) = \sum_{p \in C_j} \text{dist}(p, \mu_{C_j})^2$$

Measure for the compactness of a clustering $\mathcal{C}$:

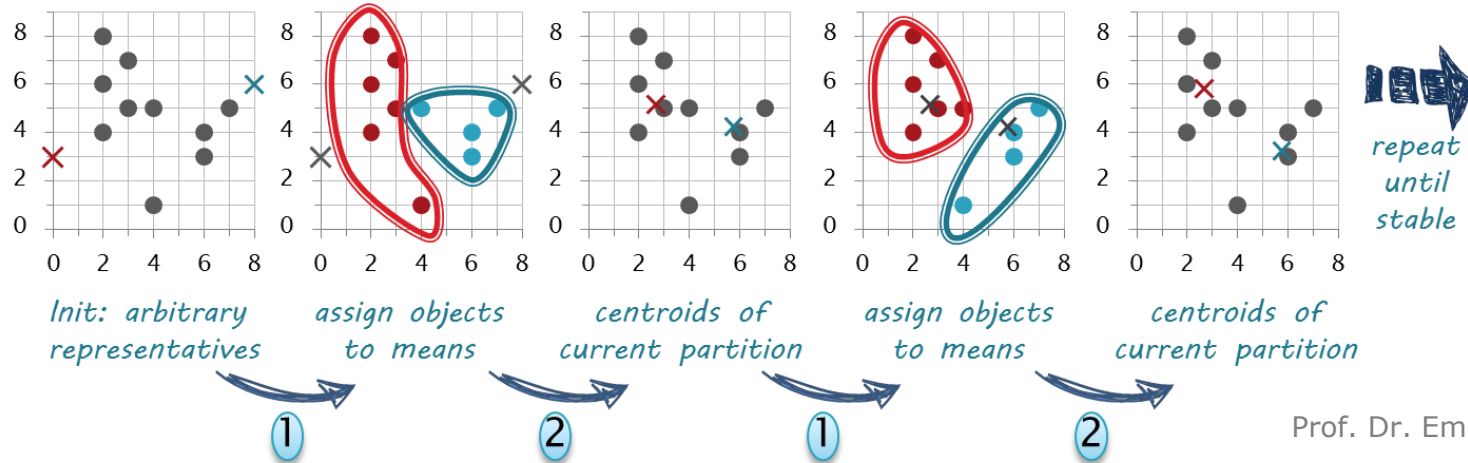
$$SSE(\mathcal{C}) = \sum_{C_j \in \mathcal{C}} SSE(C_j) = \sum_{p \in DB} \text{dist}(p, \mu_{C(p)})^2$$

Optimal Partitioning: $\underset{\mathcal{C}}{\text{argmin}} SSE(\mathcal{C})$

Optimizing the within-cluster variation is computationally challenging
(NP-hard) $\rightarrow$ use efficient heuristic algorithms

K-Means Clustering: Algorithm

- k-Means algorithm (Lloyd's algorithm) (Voronoi iterations):
- Given k, the k-means algorithm is implemented in 2 main steps:
 - Initialization: Choose k arbitrary representatives
 - Repeat until representatives do not change:
 1. Assign each object to the cluster with the nearest representative.
 2. Compute the centroids of the clusters of the current partitioning.



K-Means Clustering: Discussion

■ Strengths

- Relatively efficient: $O(tkn)$, where n = # objects, k = # clusters, and t = # iterations
- Typically: $k, t \ll n$
- Easy implementation

■ Weaknesses

- Applicable only when mean is defined
- Need to specify k , the number of clusters, in advance
- Sensitive to noisy data and outliers
- Clusters are forced to convex space partitions (Voronoi Cells)
- Result and runtime strongly depend on the initial partition; often terminates at a local optimum – however: methods for a good initialization exist

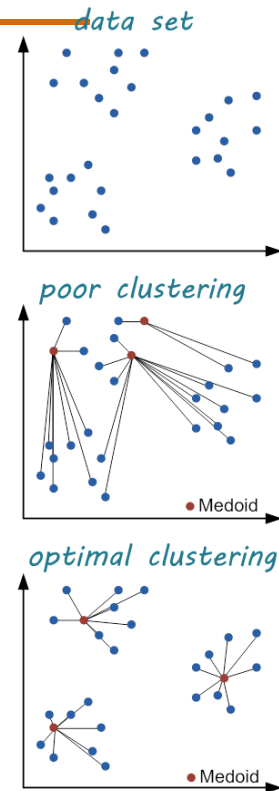
■ Several variants of the k-means method exist, e.g., ISODATA

- Extends k-means by methods to eliminate very small clusters, merging and split of clusters; however, user has to specify additional parameters

Extensions of K-Means: Basic Idea

- Problems with K-Means:
 - Applicable only when mean is defined
 - Outliers have a strong influence on the result
- The influence of outliers is intensified by the use of the *squared* error → use the absolute error (total distance instead):

$$TD(C) = \sum_{p \in C} dist(p, m_c) \text{ and } TD(\mathcal{C}) = \sum_{C_i \in \mathcal{C}} TD(C_i)$$
- Three alternatives for using the Mean as representatives:
 - **Medoid**: representative object “in the middle”
 - **Mode**: value that appears most often
 - **Median**: (artificial) representative object “in the middle”
- Objective as for k-Means: Find k representatives so that, the sum of the distances between objects and their closest representative is minimal.



K-Medoid Clustering: PAM Algorithm

Kaufman L., Rousseeuw P. J., *Finding Groups in Data: An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

PAM (Partitioning Around Medoids)

Given k , the k -medoid algorithm is implemented in 3 steps:

Initialization: Select k objects arbitrarily as initial medoids (representatives); assign each remaining (non-medoid) object to the cluster with the nearest representative, and compute TD_{current} .

Repeat

1. **For** each pair (medoid M , non-medoid N)
 - compute the value $TD_{N \leftrightarrow M}$
i.e., the value of TD for the partition that results when “swapping” M with N
2. Select the non-medoid N and medoid M for which $TD_{N \leftrightarrow M}$ is minimal
3. **If** $TD_{N \leftrightarrow M} < TD_{\text{current}}$
 - Swap N with M
 - Set $TD_{\text{current}} := TD_{N \leftrightarrow M}$

Until nothing changes

- Problem of PAM: high complexity ($O(tk(n - k)^2)$)

K-Means/Median/Mode/Medoid Clustering: Discussion

	k -Means	k -Median	K-Mode	K-Medoid
data	vector data (mean)	ordered attribute data	categorical attribute data	metric data
efficiency	high $O(tkn)$	high $O(tkn)$	high $O(tkn)$	low $O(tk(n-k)^2)$
sensitivity to outliers	high	low	low	low

■ Strength

- Easy implementation (→ many variations and optimizations in the literature)

■ Weakness

- Need to specify k , the number of clusters, in advance
- Clusters are forced to convex space partitions (Voronoi Cells)
- Result and runtime strongly depend on the initial partition; often terminates at a *local optimum* – however: methods for a good initialization exist

Initialization of Partitioning

[Fayyad, Reina, and Bradley 1998]

- Draw m different (small) samples of the dataset
- Cluster each sample to get m estimates for k representatives
 $A = (A_1, A_2, \dots, A_k)$, $B = (B_1, \dots, B_k)$, ..., $M = (M_1, \dots, M_k)$
- Then, cluster the set $DS = A \cup B \cup \dots \cup M$ m times, using the sets $A, B, \dots, M$ as respective initial partitioning
- Use the best of these m clusterings as initialization for the partitioning clustering of the whole dataset

Fayyad U., Reina C., Bradley P. S.

„Initialization of Iterative Refinement Clustering Algorithms“, *In KDD 1998*), pp. 194—198.

[Katsavounidis, Kuo, and Zhang 1994]

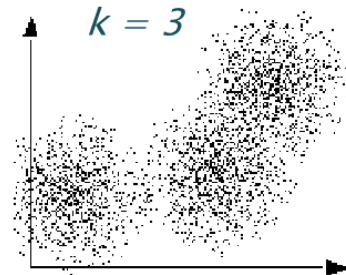
- points that are most far apart from each other are likely to belong to different clusters $\rightarrow$ use these as initial means/medoids
 - Choose the point $p_i \in DB$ with the maximum L2-norm as first centroid $\mathcal{M} = \{p_i\}$
 - Iteratively insert the point p_j into $\mathcal{M}$ with $p_j = \operatorname{argmax}_{p \in DB \setminus \mathcal{M}} \left\{ \sum_{o \in \mathcal{M}} \frac{\operatorname{dist}(p,o)}{|\mathcal{M}|} \right\}$ until $|\mathcal{M}| = k$

I. Katsavounidis, C.-C. J. Kuo, and Z. Zhang,

A new initialization technique for generalized Lloyd iteration. IEEE Signal Processing Letters, 1(10)

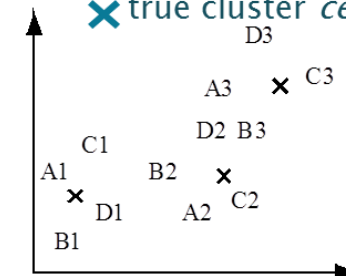
whole dataset

$k = 3$



$m = 4$ samples A, B, C, D

✕ true cluster centers



Choice of the Parameter k

Idea for a method:

- Determine a clustering for each $k = 2, \dots, n-1$
- Choose the “best” clustering

But how to measure the quality of a clustering?

- A measure should not be monotonic over k .
- The measures for the compactness of a clustering TD^2 and TD are monotonously decreasing with increasing value of k .

Silhouette-Coefficient [Kaufman & Rousseeuw 1990]

- Measure for the quality of a k -means or a k -medoid clustering that is not monotonic over k .