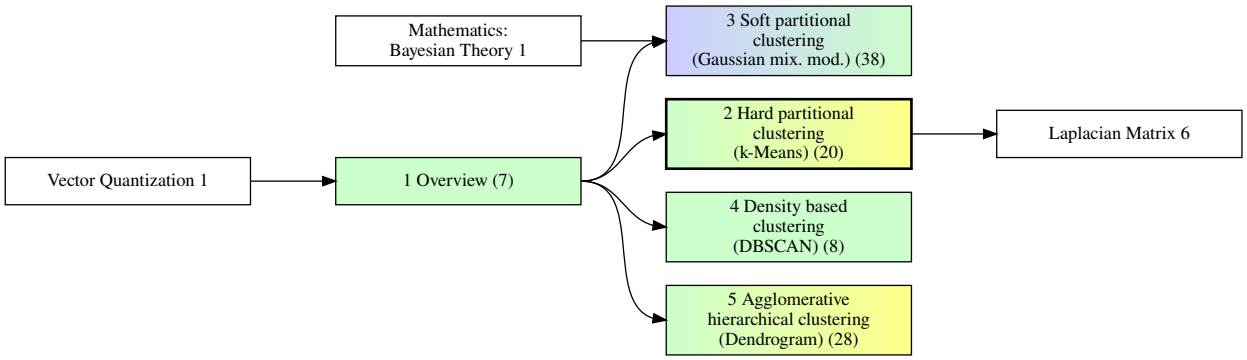
Clustering

— Lecture Notes —

Laurenz Wiskott
Institut für Neuroinformatik
Fakultät für Informatik
Ruhr-Universität Bochum, Germany, EU

22 May 2022

— Summary —



2 Hard partitional clustering considers the case of compact clusters with hard boundaries between them. Each cluster is represented by one reference vector at the mean of the corresponding data points, thus this setting is very similar to the one considered for vector quantization. The Davies-Bouldin index is introduced to measure the quality of the clustering. An application to prototypes of Northern Californian shoppers is presented.

Contents

2 Hard partitional clustering **2**

© 2009, 2011, 2014, 2016, 2020, 2022 Laurenz Wiskott (ORCID https://orcid.org/0000-0001-6237-740X, homepage https://www.ini.rub.de/PEOPLE/wiskott/). This work (except for all figures from other sources, if present) is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License, see https://creativecommons.org/licenses/by-sa/4.0/. If figures are not included for copyright reasons, they are uni colored, but the word 'Figure', 'Image', or the like in the reference is often linked to a freely available copy.
If applicable, dark red marks core text and sometimes important formulas, a filled diamond ♦ marks important formulas, and an empty diamond ◊ marks less important formulas or items that I would usually also present in a lecture to give context or dive into the details; core text and important formulas are worth remembering and learning for an exam; + marks sections that I would usually skip in a lecture.
You can also download the teaching material of this topic as zip files and then view them locally on your computer.

2.1 K-means algorithm 3
2.2 Davies-Bouldin index 4
2.3 Application: Northern Californian shoppers (→ slides) 5

These lecture notes depend on my lecture notes on vector quantization.

2 Hard partitional clustering

Learning material:[1]

☐ 16 min video 2 Hard Partitional Clustering

☐ Video List Clustering Algorithms[2] by Victor Lavrenko on YouTube

- ● 4 min video Clustering 4: Overview of k-means clustering[3] by Victor Lavrenko on YouTube
- ● 4 min video Clustering 5: The K-means algorithm[4] by Victor Lavrenko on YouTube
- ● 4 min video Clustering 6: The k-means algorithm visually[5] by Victor Lavrenko on YouTube
- ● 6 min video Clustering 7: K-means objective function[6] by Victor Lavrenko on YouTube
- ● 5 min video Clustering 8: Optimal number of clusters[7] by Victor Lavrenko on YouTube
- \+ 7 min video Clustering 9: Evaluating clustering systems[8] by Victor Lavrenko on YouTube
- \+ 13 min video Clustering 10: Intrinsic evaluation and alignment[9] by Victor Lavrenko on YouTube
- \+ 2 min video Clustering 11: Paired evaluation and Rand index[10] by Victor Lavrenko on YouTube
- \+ 9 min video Clustering 12: K-means for image representation[11] by Victor Lavrenko on YouTube
- \+ 2 min video Clustering 13: Summary[12] by Victor Lavrenko on YouTube

☐ Text below

[1] Generic instruction: Consider the (possibly nested) list of resources like a horizontal tree with an invisible root on the very left, and decide from left to right what you want to select to work through. The invisible root node has to be selected. For any selected parent node all children nodes marked with ■ or ● are mandatory and have to be selected. Children nodes marked with □ or ○ are optional and may be selected in addition to get a better understanding of the material. If a parent node has no mandatory child, then at least one optional child has to be selected. Children marked with + provide additional voluntary material that can be safely ignored, typically going beyond the scope of the section. Children of non-selected parents may be ignored. ■ and □ indicate children that cover (almost) the whole material of the section. Missing content might then be indicated by struck through references to the corresponding learning objectives. ● and ○ indicate children that do not cover (almost) the whole material of the section. They must be combined with others to be complete. Items tend to be ordered by precedence and/or recommended temporal order from top to bottom, assuming that you prefer to first watch a video before reading through lecture notes. If a detailed table of content for videos or lecture notes is given, references to learning objectives might be provided in green, 1:30 should be read as 1 min and 30 seconds, and 1'30 should be read as page 1 at about 30% of the page. Video times may be linked directly to the indicated position in the video, but be aware that the video might be downloaded anew each time you click on a time. Resources without author name are usually authored by Laurenz Wiskott and his team.

[2] https://www.youtube.com/playlist?list=PLBv09BD7ez_6lYVoZ1RzVcOPIT5Lfjo0Y, accessed 2022-04-28

[3] https://www.youtube.com/watch?v=c_oCcpN0_0s, accessed 2022-04-28

[4] https://www.youtube.com/watch?v=WZL4d47hmFs, accessed 2022-04-28

[5] https://www.youtube.com/watch?v=IJt62uaZR-M, accessed 2022-04-28

[6] https://www.youtube.com/watch?v=4s6ZTQJIMYk, accessed 2022-04-28

[7] https://www.youtube.com/watch?v=9W45yMcAOsw, accessed 2022-04-28
[8] https://www.youtube.com/watch?v=6muxKPBtTx0, accessed 2022-04-28
[9] https://www.youtube.com/watch?v=97Q3xcFitCs, accessed 2022-04-28
[10] https://www.youtube.com/watch?v=HdE9h__Xb2A, accessed 2022-04-28
[11] https://www.youtube.com/watch?v=Qx4YGU-wFjY, accessed 2022-04-28
[12] https://www.youtube.com/watch?v=lRjcVYBUYVE, accessed 2022-04-28

Quizzes and Exercises: Deepen and test your understanding with the following

- Section 2 Quiz
- Section 2 Exercises, Section 2 Solutions

2.1 K-means algorithm

A particularly **simple method** for clustering **is *K-means*,** which is identical to the LBG or generalized Lloyd algorithm we know from vector quantization, just applied to clustered data. The idea is to **represent each cluster k by a center point $\boldsymbol{c}_k$ and assign each data point $\boldsymbol{x}_n$ to one of the clusters k,** which can be written in terms of index sets $\mathcal{C}_k$. The center points and the assignment are then chosen **such that** the mean squared distance between data points and center points

$$\lozenge \qquad E := \sum_{k=1}^{K} \sum_{n \in \mathcal{C}_k} \|\boldsymbol{x}_n - \boldsymbol{c}_k\|^2 \tag{2.1}$$

is minimized. This can be interpreted, for instance, in terms of a reconstruction error. Imagine we replace each data point by its associated center point. This will lead to an error, which could be quantified by (2.1). The task is to minimize this error. There is actually a close link to *vector quantization* (D: Vektorquantisierung) here.

To achieve the minimization in practice we split the problem into two phases. First we keep the assignment fixed and optimize the position of the center points; then we keep the center points fixed and optimize the assignment.

If the assignment is fixed, it is easy to show that **the optimal choice of the center positions is** given by

$$\lozenge \qquad \boldsymbol{c}_k = \frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \boldsymbol{x}_n \,, \tag{2.2}$$

which is simple the center of gravity of the points assigned to this cluster.

If the center points are fixed, it is obvious that each point should be assigned to the nearest center position. Thus, **a *Voronoi tessallation*** (D: Dirichlet-Zerlegung) **is optimal.**

The K-means algorithm now **consists of applying these two optimizations in turn until convergence.** The initial center locations could be chosen randomly from the data points. A drawback of this and many other clustering algorithms is that **the number of clusters is not determined.** One has to decide on a proper K in advance, or one simply runs the algorithm with several different K-values and picks the best according to some criterion.

Also note that **the result of the algorithm is not necessarily a global optimum** of the error function (2.1). For instance, imagine two distinct clusters of equal size and $K = 4$. If in such a situation three center points are initialized to lie in one cluster and only one lies in the other, the algorithm will optimize this only locally with three center points in one cluster and one in the other, and it will find the better solution where two center points are in each cluster. It is therefore advisable to **run the algorithm several times** with different initial center locations **and pick the best result.**

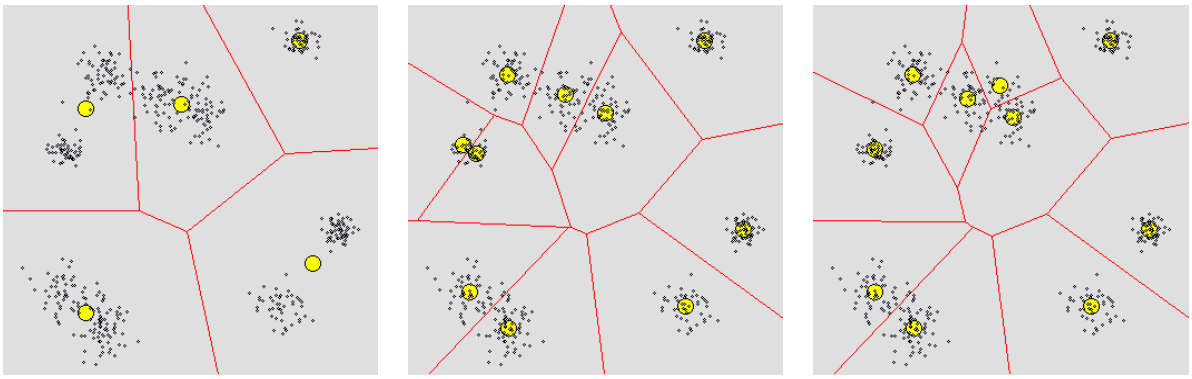


Figure 2.1: Examples of a converged K-means algorithm, once with 5 (yellow) center points (left), and two different runs with 10 center points (middle and right). The data points are drawn in black and the Voronoi tessallation in red. (Created with DemoGNG 1.5 written by Hartmut Loos and Bernd Fritzke, see https://www.demogng.de/js/demogng.html for a more recent version.)

2.2 Davies-Bouldin index

(This section is based on (Gan et al., 2007, 17.2.2).)

To evaluate the quality of a clustering a plethora of validity indices have been proposed. One of them is the ***Davies-Bouldin index*** or, for short, the DB index. **First define *cluster dispersion*** as

$$\lozenge \quad \delta_k \quad := \quad \sqrt{\frac{1}{N_k} \sum_{n \in \mathcal{C}_k} \|\boldsymbol{x}_n - \boldsymbol{c}_k\|^2} \,, \tag{2.3}$$

which can be interpreted **as a generalized standard deviation. Then define *cluster similarity*** of two clusters as

$$\lozenge \quad S_{kl} \quad := \quad \frac{\delta_k + \delta_l}{\|\boldsymbol{c}_k - \boldsymbol{c}_l\|} \,. \tag{2.4}$$

Thus, two clusters are considered similar if they have large dispersion relative to their distance.

A good clustering should be characterized by clusters being as dissimilar as possible. This should apply in particular to neighboring clusters, because it is clear that distant clusters are dissimilar in any case. Thus, **an overall validation of the clustering can be done by** the DB index

$$\lozenge \quad V_{DB} \quad := \quad \frac{1}{K} \sum_{k=1}^{K} \max_{l \neq k} S_{kl} \,. \tag{2.5}$$

The DB index does not systematically depend on K and is therefore suitable to find the best optimal number of clusters, e.g. by plotting V_{DB} and picking a pronounced minimum.

2.3 Application: Northern Californian shoppers (→ slides)

Northern Californian Shoppers

Factor	Cluster centers							Between-Cluster MSS
	Store shopaholics	Bichannel shopaholics	Time-starved worriers	Non-materialistic greens	Unwired anti-shoppers	Practical and leisure-oriented	Techno-conservatives	
Shopping enjoyment	**0.401**	**0.411**	**-0.533**	-0.148	**-0.453**	**0.407**	-0.006	L 23.75
Store enjoyment	**0.853**	**0.724**	0.237	**-0.831**	**-0.962**	-0.063	0.081	HH 70.27
Price consciousness	0.113	**0.589**	-0.119	**-0.260**	**-0.519**	**0.312**	-0.054	LL 18.35
Time consciousness	0.048	-0.061	**0.704**	0.121	-0.039	**-0.592**	**-0.399**	L 23.66
Impulse buying	**0.643**	**0.666**	**-0.286**	-0.087	**-0.761**	**-0.384**	0.238	H 39.73
Materialism	**0.797**	**0.381**	**0.447**	**-0.517**	**-0.421**	**-0.522**	**-0.297**	H 42.53
Trust	-0.058	**0.455**	**-0.493**	**0.317**	**-0.574**	0.019	**0.452**	L 24.71
Caution	**0.387**	**-0.651**	**0.636**	**-0.674**	**0.494**	0.086	**-0.376**	H 43.54
Pro-credit card	**-0.376**	**-0.360**	-0.146	**0.522**	0.156	**0.563**	**-0.445**	L 25.00
Trendsetting	0.063	**0.899**	0.142	0.003	**-0.760**	-0.111	**-0.305**	M 33.57
Pro-technology	**-0.312**	**0.733**	**0.275**	**0.475**	**-0.626**	0.142	**-0.911**	H 46.68
Pro-exercise	**-0.373**	**0.335**	-0.107	**0.405**	**-0.358**	-0.206	**0.359**	LL 16.68
Pro-environment	**-0.359**	-0.073	-0.104	**0.831**	-0.177	**-0.598**	**0.460**	M 33.83
No. (%) of observations in each cluster	149 (15.4%)	133 (13.8%)	153 (15.8%)	156 (16.1%)	136 (14.1%)	129 (13.4%)	110 (11.4%)	–

Notes: Centroid elements larger than 0.250 in magnitude are **bolded**; elements larger than 0.500 are shaded. LL, L=much below and below, respectively, the average BMSS of 34.02; M=about equal to the mean BMSS; H, HH=above, much above mean BMSS.

Table: (Mokhtarian & Ory, 2006, http://www.wctrs-society.com/... (outdated), Tab. 3)

2/4

The habbits of Northern Californian shoppers were characterized by a number of factors, then a clustering was applied and a number of prototypes identified. Factors with absolute values greater than 0.25 and greater than 0.5 are set in boldface and highlighted in green, respectively.
Table: (Mokhtarian and Ory, 2006, Tab. 3)

References

Gan, G., Ma, C., and Wu, J. (2007). *Data Clustering - Theory, Algorithms, and Applications*. ASA-SIAM Series on Statistics and Applied Probability. SIAM, Philadelphia, VA, USA.

Mokhtarian, P. L. and Ory, D. T. (2006). Shopping-related attitudes: A factor and cluster analysis of Northern California shoppers. Manuscript downloaded 2016-12-14 from `http://www.wctrs-society.com/wp/wp-content/uploads/abstracts/berkeley/D5/149/shoppingAttitudes.bergen final wctr submit.070417.doc` (outdated).