

clustering

Authors

Alexander Jung^{*}, Department of Computer Science, Aalto University, Espoo, Finland

^{*}Corresponding author: alex.jung@aalto.fi

Synonyms

segmentation

Abstract

Clustering decomposes a dataset without labels into a small number of subsets, called clusters, such that data points within the same cluster are more similar to each other than to data points in other clusters. It is a prototypical instance of unsupervised learning. Clustering methods differ in the measure of similarity between data points and in the representation of a cluster: a cluster centroid in k -means, a multivariate normal distribution in a Gaussian mixture model (GMM). They also differ in whether each data point is assigned to exactly one cluster (hard clustering) or to several with varying degrees (soft clustering). When the dataset is a graph, graph clustering partitions its nodes into densely connected subsets.

Definition

A retailer with millions of customer records has no predefined categories but still wants to group customers with similar purchasing behavior so that marketing campaigns can be targeted. Clustering addresses this type of problem: given a dataset without labels, decompose it into a small number of subsets, called clusters, such that data points within the same cluster are more similar to each other than to data points in other clusters. Clustering is a prototypical instance of unsupervised learning.

Different clustering algorithms differ in three design choices:

- 1) the measure of similarity between data points (e.g., Euclidean distance between feature vectors, edge weights in a graph);
- 2) the representation of a cluster (e.g., a cluster centroid in k -means (Lloyd, 1982), a multivariate normal distribution in a Gaussian mixture model (GMM) (Dempster et al., 1977)); and
- 3) whether each data point is assigned to exactly one cluster (hard clustering) or to several with varying degrees (soft clustering).

When the dataset is a graph rather than a set of feature vectors, graph clustering partitions the nodes into densely connected subsets with few edges between subsets (Shi and Malik, 2000; Ng et al., 2001; von Luxburg, 2007).

Fig. 1 illustrates the application of clustering to the pixels of a smartphone snapshot (Fig. 1-(a)) and to a toy graph (Fig. 1-(d)). Fig. 1-(b) shows the result of hard clustering, which partitions the pixels into disjoint regions. Fig. 1-(c) illustrates soft clustering, which assigns each pixel a vector of membership probabilities and produces graded transitions. Fig. 1-(d) depicts how graph clustering partitions a graph into densely connected subsets of nodes.

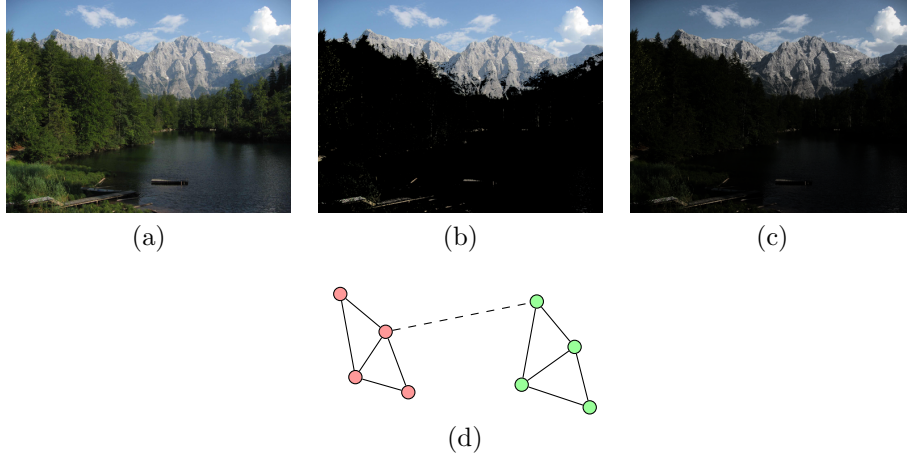


Figure 1: Flavors of clustering applied to pixels of a mountain-lake image and to a toy graph. (a) original image. (b) hard clustering via k -means with $k=2$: pixels assigned to the brightest cluster (sky/mountain) keep their original RGB value; all other pixels are blacked out. (c) soft clustering via a GMM with $k=2$: each pixel’s original RGB value is scaled by the posterior distribution probability $\mathbb{P}(\text{cluster}=c^* \mid \mathbf{x}^{(r)})$ of the same target cluster c^* ; transitions are graded rather than sharp. (d) graph clustering partitions the nodes of a graph into two subsets, such that only few edges connect nodes in different subsets. Panels (a)–(c) are produced by `pythondemos/clustering.py`

Beyond image segmentation and customer segmentation, clustering is used for document grouping in natural language processing (NLP), community detection in social networks, and as a preprocessing step to reduce a dataset to a smaller set of prototypes before supervised learning.

Cross References

cluster, cluster centroid, k -means, hard clustering, soft clustering, GMM, graph clustering, unsupervised learning

References

1. Lloyd (1982). Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2), pp. 129–137.
2. von Luxburg (2007). A tutorial on spectral clustering. *Statist. Comput.*, 17(4), pp. 395–416.
3. Ng et al. (2001). On Spectral Clustering: Analysis and an Algorithm. In: *Adv. Neural Inf. Process. Syst.*, pp. 849–856.
4. Shi and Malik (2000). Normalized Cuts and Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8), pp. 888–905.
5. Dempster et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc.: Ser. B (Methodological)*, 39(1), pp. 1–22.