cluster

Author and editors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Abstract

A cluster is a subset of the data points of a dataset that are more similar to each other than to the data points outside the subset. The measure of similarity is a design choice, such as the Euclidean distance between feature vectors or the edge weights of a similarity graph. In a hard clustering every data point belongs to exactly one cluster; in a soft clustering it belongs to every cluster with a degree of belonging, and the degrees of belonging of one data point sum to one. A clustering method represents a cluster by an object that determines the membership of any data point. Three representations are in use: a cluster centroid, as in k-means; a probability distribution, as the multivariate normal distribution of one component of a Gaussian mixture model (GMM); or a densely connected subset of the nodes of a similarity graph, as in graph clustering. The number and the shape of the clusters depend on the method and its hyperparameters: k-means delivers convex clusters, a GMM ellipsoids, and density-based spatial clustering of applications with noise (DBSCAN) any shape that a chain of nearby data points can trace.

Definition

Consider the 366 days of one year at Krems an der Donau, each a data point with a feature vector $\mathbf{x} \in \mathbb{R}^2$ that holds the morning minimum and the maximum daytime temperature of the day. A day in January, with a morning minimum near 0 and a maximum near 5 degrees, resembles the other days of January more than any day of July, whose temperatures lie some 20 degrees higher. A cluster is a subset of the data points of a dataset that are more similar to each other than to the data points outside the subset (Bishop, 2006, Sect. 9.1; Hastie et al., 2009, Sect. 14.3). Fig. 1(a) shows the two clusters that k-means finds among the days of 2024: 192 cold days and 174 warm days. Every day of January and February is a cold day, and every day of July and August is a warm day. Measured by the Euclidean distance between their feature vectors, two days of the same cluster are 8.0 degrees apart on average, a cold and a warm day 20.2 degrees, and for 364 of the 366 days the most similar other day belongs to the same cluster.

The measure of similarity between data points is a design choice (Hastie et al., 2009, Sect. 14.3). For data points with numeric feature vectors $\mathbf{x} \in \mathbb{R}^d$, two

data points are similar if the Euclidean distance between their feature vectors is small. Data points without numeric feature vectors, such as the members of a social network or the papers of a citation network, are the nodes of a similarity graph whose edge weights measure their similarity directly.

Formally, consider a dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$. A cluster is a subset $\mathcal{C}^{(c)} \subseteq \{1, \ldots, m\}$ of the indices of the data points. A hard clustering into k clusters is a partition of the indices,

$$\mathcal{C}^{(1)} \cup \cdots \cup \mathcal{C}^{(k)} = \{1, \ldots, m\} \quad \text{and} \quad \mathcal{C}^{(c)} \cap \mathcal{C}^{(c')} = \emptyset \text{ for } c \neq c', \tag{1}$$

so every data point belongs to exactly one cluster, and the hard clustering is recorded by the index of that cluster in $\{1, \ldots, k\}$ (Jain et al., 1999, Sect. 2). The two clusters of Fig. 1(a) form a partition (1) of the 366 days. A soft clustering replaces the index by a degree of belonging $u_c^{(r)}$ of data point r to every cluster $c = 1, \ldots, k$,

$$u_c^{(r)} \geq 0 \quad \text{and} \quad \sum_{c=1}^{k} u_c^{(r)} = 1 \quad \text{for } r = 1, \ldots, m, \tag{2}$$

so the degrees of belonging of one data point form a vector in the probability simplex Δ^{k-1}. A cluster is then no longer a subset but the map $r \mapsto u_c^{(r)}$ that grades the membership of every data point. A hard clustering is the special case where every degree of belonging is 0 or 1: the cluster $\mathcal{C}^{(c)}$ then consists of the data points with $u_c^{(r)} = 1$, and these clusters form a partition (1).

A clustering method does not store a cluster as a list of indices. It represents the cluster by an object from which the membership of any data point follows, including a data point that does not belong to the dataset. This object is the hypothesis that the method learns, and generalization to new data points is what it is for. Three representations are in use (Fig. 1).

First, a cluster centroid $\mathbf{w}^{(c)} \in \mathbb{R}^d$ is a single point of the feature space (Bishop, 2006, Sect. 9.1). It represents a cluster as the set of data points that are nearer to it than to any other cluster centroid,

$$\mathcal{C}^{(c)} = \left\{ r : \left\| \mathbf{x}^{(r)} - \mathbf{w}^{(c)} \right\|_2 \leq \left\| \mathbf{x}^{(r)} - \mathbf{w}^{(c')} \right\|_2 \text{ for all } c' \neq c \right\}. \tag{3}$$

This is the hard clustering delivered by k-means, whose cluster centroids minimize the clustering error, an instance of empirical risk minimization (ERM). The region of the feature space that satisfies the inequalities in the cluster (3) is an intersection of halfspaces, one for each other cluster centroid, and therefore convex. For the days at Krems and $k = 2$, the two cluster centroids are $(1.5, 10.5)$ and $(13.4, 26.1)$, and the boundary between the two clusters is the straight line through their midpoint at a right angle to their connecting segment (Fig. 1(a)).

Second, a probability distribution represents a cluster. A Gaussian mixture model (GMM) represents cluster c by a multivariate normal distribution with mean $\boldsymbol{\mu}^{(c)}$ and covariance matrix $\mathbf{C}^{(c)}$, drawn with probability p_c. The degree of

belonging of a data point to cluster c is the posterior distribution of its cluster index given its feature vector, evaluated at c (Bishop, 2006, Sect. 9.2),

$$u_c^{(r)} = \frac{p_c \, \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right)}{\sum_{c'=1}^{k} p_{c'} \, \mathcal{N}\left(\boldsymbol{\mu}^{(c')}, \mathbf{C}^{(c')}\right)\left(\mathbf{x}^{(r)}\right)}. \tag{4}$$

The degrees of belonging (4) satisfy the constraints (2), so the result is a soft clustering. The level sets of the probability density function (pdf) of a multivariate normal distribution are ellipsoids whose axes are the eigenvectors of the covariance matrix (Bishop, 2006, Sect. 2.3), so a cluster represented this way is an ellipsoid rather than an intersection of halfspaces. For the days at Krems, the covariance matrices learned by the expectation–maximization (EM) algorithm have a long axis 2.8 and 2.0 times the short one, tilted by 50 and 46 degrees, so each cluster is elongated along the direction in which both temperatures rise together (Fig. 1(b)). The larger of the two degrees of belonging names the k-means cluster for 335 of the 366 days, and 58 days have a degree of belonging between 0.2 and 0.8 to both clusters. Fig. 1(d) contrasts the two assignments over the year: the k-means index jumps between 0 and 1, while the degree of belonging to the warm cluster is 0 on every day from November to February and passes through intermediate values on days from March to October.

Third, for data points that are the nodes $\mathcal{V} = \{1, \ldots, m\}$ of a similarity graph with edge weights $A_{r,r'} \geq 0$, a cluster is a subset $\mathcal{C} \subseteq \mathcal{V}$ of nodes that are densely connected among themselves and sparsely connected to the rest. Graph clustering methods measure this by the sum of the edge weights of the edges leaving the subset, divided by its size,

$$\frac{1}{|\mathcal{C}|} \sum_{r \in \mathcal{C}} \sum_{r' \notin \mathcal{C}} A_{r,r'}, \tag{5}$$

and a hard clustering into k clusters is good if the sum of the ratios (5) over its clusters, the ratio cut, is small (von Luxburg, 2007, Sect. 5, Eq. (1)). Spectral clustering minimizes the ratio cut approximately. For the days at Krems, the k-nearest-neighbor graph (k-NN graph) that joins each day to its 10 nearest neighbors, with unit edge weights, has 2209 edges; 2174 of them join two days of the same k-means cluster and 35 join a cold with a warm day. The ratio cut of the two clusters is 0.38, against 12.07 for the split of the days into those with an odd and those with an even index (Fig. 1(c)).

Which subsets of a dataset are clusters therefore depends on the method and its hyperparameters, not on the dataset alone. The number k of clusters is an input to k-means and to a GMM, while density-based spatial clustering of applications with noise (DBSCAN) determines it from a distance threshold and a minimum cluster size. The shape of a cluster follows from its representation: an intersection of halfspaces for k-means, an ellipsoid for a GMM, and any shape that a chain of nearby data points can trace for DBSCAN and for graph clustering. The same 366 days are split by a straight line in Fig. 1(a), while in Fig. 1(b) each day receives a degree of belonging to each of two tilted ellipses.

The choice of representation is a choice of hypothesis space, and the number of clusters is chosen like any other hyperparameter, since the clustering error alone cannot choose it.

Clusters serve as categories where no label provides them, the setting of unsupervised learning. An energy supplier can treat the cold days of Fig. 1(a) as the heating season and classify a new day by the side of the boundary line on which its feature vector falls. In image segmentation, a cluster of pixels with similar colors is one region of the image (Bishop, 2006, Sect. 9.1.1), and in customer segmentation a cluster of customers with similar purchase records is one market segment (Wedel and Kamakura, 2000).

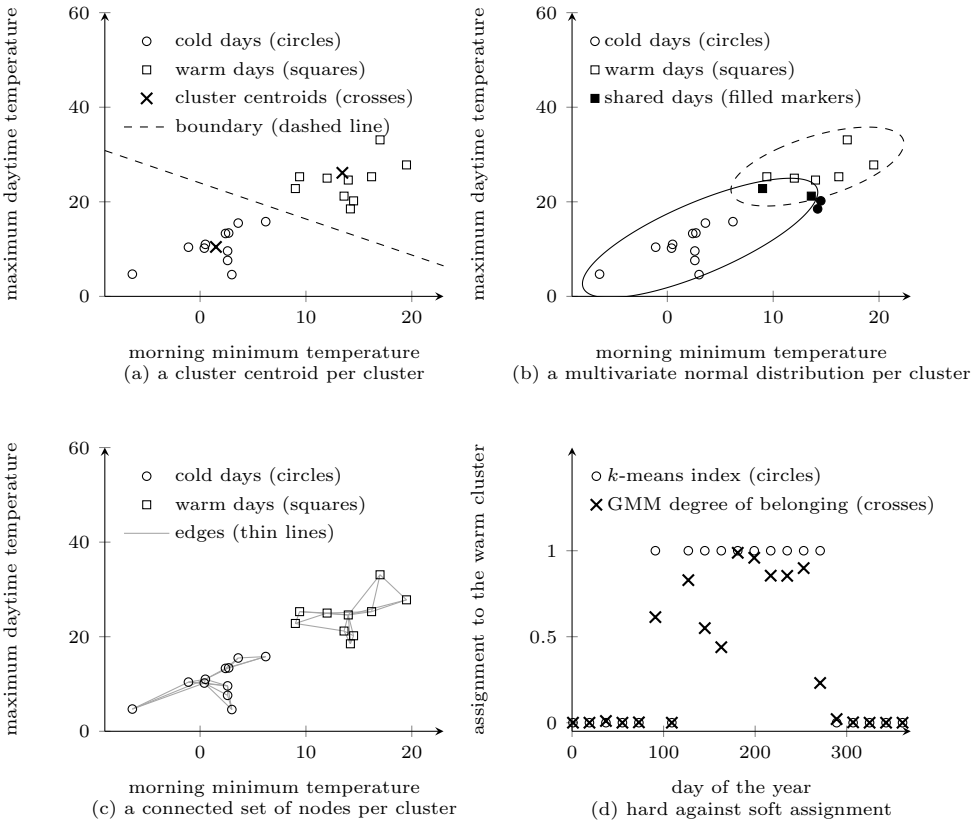


Figure 1: Every 18th of the 366 days of 2024 at Krems, 21 days, under the three representations of a cluster. (a) The days drawn as cold days (circles) or warm days (squares) by the nearer of the two cluster centroids of k-means (crosses); the boundary between the two clusters is a straight line (dashed line). (b) The two multivariate normal distributions of a GMM, the cold cluster (solid ellipse) and the warm cluster (dashed ellipse), each drawn as the level set of its pdf at twice the standard deviation along each axis; a day is drawn by the cluster with the larger degree of belonging, and filled where that degree of belonging is below 0.8 (filled markers). (c) The 3-nearest-neighbor graph of the 21 drawn days, which joins each day to the three days most similar to it (thin lines); all 41 edges join two days of the same cluster. (d) The same 21 days by their day of the year: the k-means assignment, 0 for a cold and 1 for a warm day (circles), and the degree of belonging to the warm cluster under the GMM (crosses). Data generated by `pythondemos/cluster.py`

Cross References

clustering, cluster centroid, clustering error, hard clustering, soft clustering, degree of belonging, k-means, GMM, multivariate normal distribution, graph

clustering, similarity graph, k-NN graph, edge weight, spectral clustering, DBSCAN, data point, feature vector, Euclidean distance, feature space, hyperparameter, convex, probability simplex, unsupervised learning

References

1. Bishop (2006). Pattern Recognition and Machine Learning. Springer Science+Business Media, New York, NY, USA.

2. Jain et al. (1999). Data clustering: a review. ACM Computing Surveys, 31(3), pp. 264–323.

3. von Luxburg (2007). A tutorial on spectral clustering. Statist. Comput., 17(4), pp. 395–416.

4. Wedel and Kamakura (2000). Market Segmentation: Conceptual and Methodological Foundations. 2nd ed., Kluwer Academic Publishers, Boston, MA.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.