cluster centroid

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 centroid is a vector that represents a cluster of data points: among all vectors, it minimizes the summed loss over the data points of the cluster, where the loss measures how poorly a vector stands in for a data point. Under the squared Euclidean distance the cluster centroid is the mean of the feature vectors in the cluster, under the Euclidean distance their geometric median (GM), and under the sum of absolute differences the median of each feature. k-means assigns a data point to its nearest cluster centroid, a rule that is a linear classifier whose decision boundary between two cluster centroids is the hyperplane through their midpoint and perpendicular to their difference. The averaging step of Lloyd's algorithm recomputes each cluster centroid as the mean of its cluster, and at a fixed point every cluster centroid is the mean of the data points nearest to it. The mean moves with every data point, so one outlier shifts it by the displacement divided by the size of the cluster, while the GM and the coordinate-wise median move little. A soft clustering method, such as a Gaussian mixture model (GMM), uses the mean weighted by the degrees of belonging as its cluster centroid. Replacing every data point by its cluster centroid compresses a dataset to k vectors and one cluster index per data point.

Definition

Consider the 366 days of 2024 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. k-means with $k = 2$ decomposes these days into 192 cold days and 174 warm days, and it describes each cluster by a single vector: a typical cold day, $(1.5, 10.5)$, and a typical warm day, $(13.4, 26.1)$. Such a vector is a cluster centroid. A hard clustering method that represents each cluster $c = 1, \ldots, k$ by a cluster centroid $\mathbf{w}^{(c)} \in \mathbb{R}^d$ assigns a data point to the cluster whose cluster centroid is nearest (Bishop, 2006, Sect. 9.1; Hastie et al., 2009, Sect. 14.3.6), as Fig. 1(a) shows for the days at Krems.

Which vector represents a cluster best is decided by a loss function $L(\mathbf{x}, \mathbf{w})$ that measures how poorly a vector $\mathbf{w}$ stands in for the feature vector $\mathbf{x}$ of a data point. The cluster centroid of a cluster with index set $\mathcal{C}^{(c)} \subseteq \{1, \ldots, m\}$ is

a vector that minimizes the summed loss over its data points,

$$\mathbf{w}^{(c)} \in \operatorname*{arg\,min}_{\mathbf{w}\in\mathbb{R}^d} \sum_{r\in\mathcal{C}^{(c)}} L\left(\mathbf{x}^{(r)}, \mathbf{w}\right). \tag{1}$$

For the squared Euclidean distance $L(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|_2^2$, the summed loss is a strictly convex quadratic function of $\mathbf{w}$, and its unique minimizer is the mean of the feature vectors,

$$\mathbf{w}^{(c)} = \frac{1}{\left|\mathcal{C}^{(c)}\right|} \sum_{r\in\mathcal{C}^{(c)}} \mathbf{x}^{(r)}, \tag{2}$$

since the gradient of the summed loss vanishes exactly where the deviations $\mathbf{x}^{(r)} - \mathbf{w}$ sum to zero (MacQueen, 1967). The two cluster centroids of the days at Krems are the means (2) of the cold and of the warm days. Under the Euclidean distance itself, $L(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|_2$, the minimizer of the summed loss (1) is the geometric median (GM) of the feature vectors (Boyd and Vandenberghe, 2004, Sect. 8.7.1). It has no closed form and is computed by Weiszfeld's fixed-point iteration (Weiszfeld, 1937),

$$\mathbf{w}^{(t+1)} = \mathcal{F}\big(\mathbf{w}^{(t)}\big) := \frac{\sum_{r\in\mathcal{C}^{(c)}} \mathbf{x}^{(r)} \big/ \big\|\mathbf{x}^{(r)} - \mathbf{w}^{(t)}\big\|_2}{\sum_{r\in\mathcal{C}^{(c)}} 1 \big/ \big\|\mathbf{x}^{(r)} - \mathbf{w}^{(t)}\big\|_2} \quad \text{for } t = 0, 1, \ldots, \tag{3}$$

a mean of the feature vectors weighted by the reciprocal of their distance to the current iterate. A fixed point of the iteration (3) that coincides with no feature vector is a point where the unit vectors towards the feature vectors sum to zero, which is the optimality condition of the GM, and the summed loss never increases from one iterate to the next (Weiszfeld, 1937). Under the sum of absolute differences, $L(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|_1$, the summed loss separates into one problem per feature, so the cluster centroid holds the median of each feature over the cluster. For the warm days at Krems, the three choices give $(13.4, 26.1)$, $(13.5, 26.0)$, and $(13.7, 26.1)$, within 0.3 degrees of each other; Weiszfeld's iteration started from the mean stops after 60 iterations.

For a fixed cluster, the minimization (1) is empirical risk minimization (ERM) over the hypothesis space of constant maps $h(\mathbf{x}) = \mathbf{w}$. The cluster centroid is the hypothesis that predicts one vector for every data point of the cluster, and the summed loss is its training error up to the factor $1/|\mathcal{C}^{(c)}|$. The k cluster centroids together are the model parameters of the hypothesis $h(\mathbf{x}) = \mathbf{w}^{(c^\star(\mathbf{x}))}$ learned by k-means, which sends a feature vector to its nearest cluster centroid,

$$c^\star(\mathbf{x}) := \operatorname*{arg\,min}_{c=1,\ldots,k} \left\|\mathbf{x} - \mathbf{w}^{(c)}\right\|_2. \tag{4}$$

The average of the loss $\|\mathbf{x} - h(\mathbf{x})\|_2^2$ over the dataset is the clustering error. The nearest-centroid rule (4) is a linear classifier. For two cluster centroids, a feature vector is nearer to $\mathbf{w}^{(1)}$ than to $\mathbf{w}^{(2)}$ if and only if

$$\big(\mathbf{w}^{(2)} - \mathbf{w}^{(1)}\big)^\top \mathbf{x} \le \frac{1}{2}\Big(\left\|\mathbf{w}^{(2)}\right\|_2^2 - \left\|\mathbf{w}^{(1)}\right\|_2^2\Big), \tag{5}$$

which follows from expanding the two squared Euclidean distances. The decision boundary of the linear rule (5) is the hyperplane through the midpoint of the two cluster centroids and perpendicular to their difference (Bishop, 2006, Sect. 9.1), the dashed line of Fig. 1(a). Each cluster is an intersection of halfspaces and so a convex region. The rule (4) also assigns a data point outside the dataset. A day with morning minimum 5 and maximum 18 degrees lies at distances 8.3 and 11.7 from the two cluster centroids of the days at Krems, so it is a cold day. The same rule serves classification when each class of a labeled dataset is represented by one or several cluster centroids and a new data point receives the label of the nearest one (Hastie et al., 2009, Sect. 13.2.1).

Lloyd's algorithm for k-means alternates the assignment (4) of every data point with an averaging step that recomputes each cluster centroid as the mean (2) of the data points assigned to it (Lloyd, 1982). At a fixed point, each cluster centroid is the mean of the data points nearest to it. Two conditions then hold at the same time: each cluster centroid minimizes the summed loss (1) over its cluster, and each cluster consists of the data points nearest to its cluster centroid (Hastie et al., 2009, Sect. 14.3.6). For the days at Krems, started from the means of the colder and of the warmer half of the days, Lloyd's algorithm reaches this fixed point after 2 iterations. Changing the loss function changes only the averaging step: the GM gives k-medians, and restricting the cluster centroid to be one of the data points of its cluster gives k-medoids, which needs only the pairwise distances between data points (Hastie et al., 2009, Sect. 14.3.10; Park and Jun, 2009; Shalev-Shwartz and Ben-David, 2014, Sect. 22.2). For a Bregman divergence as loss function, the mean (2) remains the cluster centroid (Banerjee et al., 2005, Prop. 1).

The mean (2) moves with every data point of its cluster: adding a vector to one feature vector adds the same vector divided by $|\mathcal{C}^{(c)}|$ to the mean, whatever the other data points are. If the maximum temperature of the hottest warm day at Krems were recorded as 999.9, a value some archives use to mark a missing record, the mean of the 174 warm days would move by $965.6/174 = 5.5$ degrees (Fig. 1(b)). The GM would move by 0.03 degrees and the coordinate-wise median not at all. The squared Euclidean distance gives the largest distances the largest weight, which is why cluster centroids built from an unsquared norm are less sensitive to outliers (Hastie et al., 2009, Sect. 14.3.10).

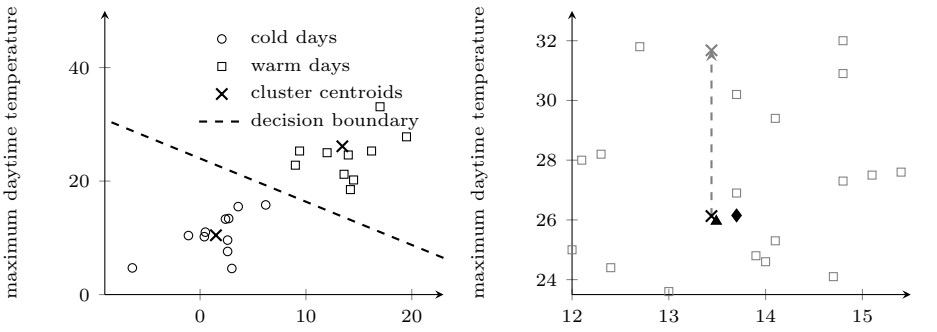
A soft clustering method assigns each data point r to every cluster c with a degree of belonging $u_c^{(r)} \in [0, 1]$, with $\sum_{c=1}^{k} u_c^{(r)} = 1$. Its cluster centroid is the mean of all feature vectors weighted by their degrees of belonging,

$$\boldsymbol{\mu}^{(c)} = \frac{\sum_{r=1}^{m} u_c^{(r)} \mathbf{x}^{(r)}}{\sum_{r=1}^{m} u_c^{(r)}}, \tag{6}$$

whose denominator is the effective number of data points in the cluster. For a Gaussian mixture model (GMM), which represents the cluster c by a multivariate normal distribution with mean $\boldsymbol{\mu}^{(c)}$ and covariance matrix $\mathbf{C}^{(c)}$, the degree of belonging is the posterior distribution of the cluster index, and the weighted mean (6) is the M-step of the expectation–maximization (EM) algorithm for $\boldsymbol{\mu}^{(c)}$

(Bishop, 2006, Sect. 9.2.2). Hard clustering is the case $u_c^{(r)} \in \{0, 1\}$, where the weighted mean (6) is the mean (2). For the days at Krems, the EM algorithm started from the two k-means cluster centroids stops at the means $(3.1, 12.4)$ and $(14.2, 27.5)$, which lie 2.5 and 1.5 degrees from them. Of the 366 days, 58 have no degree of belonging above 0.8, and each of them is counted in part by both cluster centroids.

Replacing every data point by its cluster centroid compresses a dataset to k vectors and one cluster index per data point, which is vector quantization. The 366 days at Krems become two typical days and one bit per day, and a photograph whose pixels are clustered by color becomes a palette of k colors and one palette index per pixel (Bishop, 2006, Sect. 9.1.1; Shalev-Shwartz and Ben-David, 2014, Sect. 22.2). The clustering error is the average distortion of this compression, and minimizing it over the cluster centroids is the task from which Lloyd's algorithm arose (Lloyd, 1982).



(a) 21 of the 366 days, two cluster centroids

(b) the cluster centroid of the warm days

Figure 1: (a) Every 18th of the 366 days of 2024 at Krems, 21 days, drawn as cold days (circles) or warm days (squares) by the nearer of the two cluster centroids (crosses), the means of the cold and of the warm days; the dashed line is the decision boundary of the nearest-centroid rule, the hyperplane through the midpoint of the two cluster centroids and perpendicular to their difference. (b) The 18 of the 174 warm days that fall inside the window (gray squares) and the cluster centroid of all 174 warm days under three losses: their mean (black cross) under the squared Euclidean distance, their GM (filled triangle) under the Euclidean distance, and their coordinate-wise median (filled diamond) under the sum of absolute differences; the dashed arrow leads from the mean to the mean after the maximum temperature of one warm day is replaced by 999.9 (gray cross), a move of 5.5 degrees, while the two medians move by 0.03 degrees and not at all. Data generated by `pythondemos/clustercentroid.py`

Cross References

clustering, cluster, clustering error, k-means, Lloyd's algorithm, k-means++, hard clustering, soft clustering, degree of belonging, GMM, EM, mean, median, GM, Euclidean distance, norm, outlier, linear classifier, decision boundary, hyperplane, fixed-point iteration, ERM, hypothesis, model parameter, Bregman divergence, feature vector

References

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

2. Boyd and Vandenberghe (2004). Convex Optimization. Cambridge Univ. Press, Cambridge, U.K.

3. Banerjee et al. (2005). Clustering with Bregman Divergences. J. Mach. Learn. Res., 6(58), pp. 1705–1749.

4. Lloyd (1982). Least squares quantization in PCM. IEEE Trans. Inf. Theory, 28(2), pp. 129–137.

5. MacQueen (1967). Some methods for classification and analysis of multivariate observations. In: Proc. Fifth Berkeley Symp. Math. Statist. Probab., pp. 281–297. University of California Press.

6. Park and Jun (2009). A simple and fast algorithm for K-medoids clustering. Expert Syst. Appl., 36(2, Part 2), pp. 3336–3341.

7. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

8. Weiszfeld (1937). Sur le point pour lequel la somme des distances de n points donnés est minimum. Tôhoku Math. J., 43, pp. 355–386.

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