clustering error

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

The clustering error of a clustering is the average, over the data points of a dataset, of a loss that measures how poorly a data point is represented by the cluster it is assigned to. For hard clustering methods that represent a cluster by a cluster centroid, such as k-means, the loss is the squared Euclidean distance to the nearest cluster centroid. The Euclidean distance itself or the sum of absolute differences gives a variant whose best cluster centroids are medians instead of means and which is less sensitive to outliers. Soft clustering with a Gaussian mixture model (GMM) uses the negative logarithm of the probability density function (pdf) as the loss, so its clustering error is the negative log-likelihood function per data point. In every case the clustering error is the training error of a hypothesis, and minimizing it is empirical risk minimization (ERM). Lloyd's algorithm and the expectation–maximization (EM) algorithm never increase it but may stop at a local minimum. The clustering error is non-increasing in the number of clusters and vanishes when every data point is its own cluster, so it cannot choose that number.

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 clustering method decomposes these days into a few clusters, and clustering methods differ in how they represent a cluster. One representation is a cluster centroid, a vector $\mathbf{w}^{(c)} \in \mathbb{R}^d$ for each cluster $c = 1, \ldots, k$; each data point is assigned to the cluster with the nearest cluster centroid, which is the hard clustering delivered by k-means. A natural measure of how well the cluster centroids represent the dataset $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ is the average, over the data points, of the minimum squared Euclidean distance to the cluster centroids (Bishop, 2006, Sect. 9.1, Eq. (9.1)),

$$\frac{1}{m} \sum_{r=1}^{m} \min_{c=1,\ldots,k} \left\| \mathbf{x}^{(r)} - \mathbf{w}^{(c)} \right\|_2^2. \tag{1}$$

The clustering error of k-means is the average (1). For two cluster centroids and the days at Krems, it is 42.0; Fig. 1(a) draws the distances it averages for some of the days.

In the end, the clustering error (1) is the average of a loss, for a special choice of the loss function. The hypothesis learned by k-means is the map $h(\mathbf{x}) = \mathbf{w}^{(c^\star(\mathbf{x}))}$ that sends a feature vector to its nearest cluster centroid, with $c^\star(\mathbf{x}) = \arg\min_{c=1,\ldots,k} \left\|\mathbf{x} - \mathbf{w}^{(c)}\right\|_2$. The loss of a data point is (Shalev-Shwartz and Ben-David, 2014, Sect. 22.2)

$$L\left(\mathbf{x}, h\right) := \left\|\mathbf{x} - h(\mathbf{x})\right\|_2^2 = \min_{c=1,\ldots,k} \left\|\mathbf{x} - \mathbf{w}^{(c)}\right\|_2^2. \tag{2}$$

A data point carries no label here: the loss (2) compares the feature vector with the cluster centroid that stands in for it. The clustering error is the average of the loss (2) over the dataset, so it is the training error of the hypothesis h, and choosing cluster centroids that minimize it is empirical risk minimization (ERM) over the hypothesis space of all maps of this form.

The squared Euclidean distance in the loss (2) is one choice among several. Replacing the squared norm by the norm itself, $L\left(\mathbf{x}, h\right) = \left\|\mathbf{x} - h(\mathbf{x})\right\|_2$, or by the ℓ_1-norm $\left\|\mathbf{x} - h(\mathbf{x})\right\|_1$, the sum of the absolute differences of the features, keeps the assignment of a data point to its nearest cluster centroid. It changes which cluster centroid is best for a given cluster. Under the squared Euclidean distance, it is the mean of the data points in the cluster, under the Euclidean distance their geometric median (GM), computed by Weiszfeld's fixed-point iteration (Weiszfeld, 1937), and under the ℓ_1-norm the median of each feature separately. A single far-away data point, an outlier, moves the mean of its cluster in proportion to its distance but moves either median little, so the clustering errors built from an unsquared norm are less sensitive to outliers (Hastie et al., 2009, Sect. 14.3.10). For the days at Krems and two cluster centroids, the three choices assign more than 99 percent of the days to the same cluster. Their clustering errors, 42.0, 5.8, and 7.6, are not comparable, since the first is measured in squared degrees and the other two in degrees. The squared Euclidean distance can also be replaced by a Bregman divergence, for which the mean remains the best cluster centroid (Banerjee et al., 2005).

Another choice is used by soft clustering methods, which represent a cluster by a probability distribution rather than by a single vector and assign each data point to every cluster with a degree of belonging. A Gaussian mixture model (GMM) represents a clustering by its model parameters: the cluster c is a multivariate normal distribution with mean $\boldsymbol{\mu}^{(c)}$ and covariance matrix $\mathbf{C}^{(c)}$, chosen with probability p_c. The clustering error of these model parameters is the negative log-likelihood function of the dataset per data point (Bishop, 2006, Sect. 9.2, Eq. (9.14)),

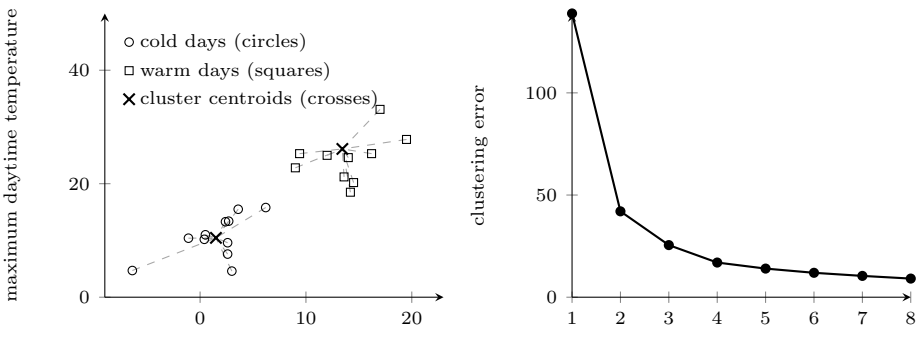
$$-\frac{1}{m} \sum_{r=1}^{m} \ln \sum_{c=1}^{k} p_c \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right), \tag{3}$$

again the average of a loss, the negative logarithm of the probability density function (pdf) at the feature vector. Minimizing the clustering error (3) is maximum likelihood estimation of the model parameters (Bishop, 2006, Sect. 9.2).

The k-means clustering error is a limit of this choice: for equal probabilities $p_c = 1/k$ and covariance matrices $\mathbf{C}^{(c)} = \sigma^2\mathbf{I}$, the loss of a data point in the clustering error (3) is, as $\sigma \to 0$ and up to the factor $1/(2\sigma^2)$ and an additive constant, the loss (2) (Bishop, 2006, Sect. 9.3.2).

Each of these clustering errors is a non-convex function of the cluster centroids or model parameters: with a single data point at $x = 0$ and two scalar cluster centroids $w^{(1)}, w^{(2)} \in \mathbb{R}$, the clustering error (1) is $\min\{(w^{(1)})^2, (w^{(2)})^2\}$, which vanishes at $(1, 0)$ and at $(0, 1)$ but not at their midpoint. Lloyd's algorithm for k-means (Lloyd, 1982) and the expectation–maximization (EM) algorithm for a GMM are fixed-point iterations that never increase the clustering error and stop at a fixed point, which can be a local minimum rather than a global one (Bishop, 2006, Sect. 9.1). They are therefore run from several initializations, keeping the run with the smallest clustering error (Hastie et al., 2009, Sect. 14.3.6). For the days at Krems, Lloyd's algorithm started from the means of the colder and of the warmer half of the days lowers the clustering error from 43.0 to 42.0 in two iterations. At the fixed point, 192 cold days share the cluster centroid $(1.5, 10.5)$ and 174 warm days the cluster centroid $(13.4, 26.1)$; every day of January and February is a cold day, and every day of July and August a warm day.

The clustering error cannot choose the number k of clusters. An additional cluster centroid placed on any data point cannot increase it, so the smallest clustering error attainable with k cluster centroids is non-increasing in k, and it is zero at $k = m$, where every data point is its own cluster centroid. This is the training error of model selection in another guise: a larger hypothesis space never has a larger training error. Fig. 1(b) shows the curve for the days at Krems. With one cluster centroid the clustering error is the sum of the variances of the two features, 138.9; the second cluster centroid removes 96.8 of it, more than the next six cluster centroids together (32.9). Such a kink in the curve is the usual informal guide to k (Hastie et al., 2009, Sect. 14.3.11); otherwise k is a hyperparameter, chosen by a criterion that penalizes it or by the performance of a prediction task that uses the clusters.



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

(b) smallest clustering error found

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 two cluster centroids (crosses); the clustering error (1) averages the squared lengths of the dashed segments over all 366 days. (b) The smallest clustering error found with $k = 1, \ldots, 8$ cluster centroids: it falls with k and would reach zero at $k = 366$. Data generated by `pythondemos/clusteringerror.py`

Cross References

clustering, cluster, cluster centroid, hard clustering, soft clustering, degree of belonging, k-means, Lloyd's algorithm, GMM, EM, maximum likelihood, loss, loss function, training error, ERM, Euclidean distance, norm, GM, median, Bregman divergence, outlier, hyperparameter, model selection

References

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

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

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

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

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

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