soft clustering

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

Synonyms

fuzzy clustering

Abstract

Soft clustering assigns each data point to every cluster with a degree of belonging, a number between zero and one, so that the degrees of belonging of one data point form a vector in the probability simplex. Hard clustering is the special case in which this vector is a vertex of the simplex, with one entry equal to one and the others equal to zero. For numeric feature vectors, the principled choice of the degrees of belonging is the posterior distribution of the cluster index under a Gaussian mixture model (GMM) whose model parameters are learned by the expectation–maximization (EM) algorithm. With equal cluster probabilities and covariance matrices $\sigma^2\mathbf{I}$, the degrees of belonging tend to the vertices as σ^2 shrinks, and the EM algorithm becomes Lloyd's algorithm for k-means. Fuzzy c-means delivers degrees of belonging without a probabilistic model by minimizing a weighted clustering error over the simplex, and it also reduces to k-means as its fuzzifier tends to one. Assigning each data point to the cluster with the largest degree of belonging turns a soft clustering into a hard clustering, and one minus that largest degree of belonging measures how far the assignment is from certain.

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 $k = 2$ clusters assigns every day to the nearer of two cluster centroids, $(1.5, 10.5)$ for 192 cold days and $(13.4, 26.1)$ for 174 warm days. A day in April whose temperatures lie halfway between the two cluster centroids becomes a cold day or a warm day all the same, and nothing records that it was a borderline case. Soft clustering keeps that information: it assigns each data point to every cluster with a degree of belonging $u_c^{(r)}$, a number between 0 and 1 for data point $r = 1, \ldots, m$ and cluster $c = 1, \ldots, k$ (Bishop, 2006, Sect. 9.3.2; Murphy, 2012, Sect. 11.2.3). The

degrees of belonging of one data point form the vector

$$\mathbf{u}^{(r)} := \left(u_1^{(r)}, \ldots, u_k^{(r)}\right)^\top \in \Delta^{k-1}, \quad \text{i.e., } u_c^{(r)} \geq 0 \text{ and } \sum_{c=1}^{k} u_c^{(r)} = 1, \tag{1}$$

a point of the probability simplex Δ^{k-1}. A hard clustering is the special case in which every vector (1) is a vertex of the simplex: one entry is 1 and the other $k-1$ entries are 0. A soft clustering method learns a hypothesis $h : \mathcal{X} \to \Delta^{k-1}$ that maps a feature vector to a degree of belonging vector, so it can be evaluated for a day that did not belong to the dataset. A hard clustering method learns a map into the k vertices only.

For data points with numeric feature vectors, the principled choice of the degrees of belonging is a probabilistic model, the Gaussian mixture model (GMM). It generates a data point by drawing a cluster index I with probabilities $p_1, \ldots, p_k$ and then the feature vector from the multivariate normal distribution of that cluster, with mean $\boldsymbol{\mu}^{(c)}$ and covariance matrix $\mathbf{C}^{(c)}$. The degree of belonging of data point $\mathbf{x}^{(r)}$ to cluster c is the posterior distribution of its cluster index,

$$u_c^{(r)} := \mathbb{P}\left(I = c \mid \mathbf{x}^{(r)}\right) = \frac{p_c \, \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right)}{\sum_{c'=1}^{k} p_{c'} \, \mathcal{N}\left(\boldsymbol{\mu}^{(c')}, \mathbf{C}^{(c')}\right)\left(\mathbf{x}^{(r)}\right)}, \tag{2}$$

which Bishop (2006), Sect. 9.2 calls the responsibility of cluster c for the data point. The model parameters are learned by maximum likelihood with the expectation–maximization (EM) algorithm, whose E-step computes the degrees of belonging (2) and whose M-step recomputes the model parameters as averages weighted by them (Bishop, 2006, Sect. 9.2.2). Maximum likelihood is empirical risk minimization (ERM) with the negative logarithm of the probability density function (pdf) as the loss, so the learned model parameters minimize a clustering error, and the EM algorithm is a fixed-point iteration that never increases it (Dempster et al., 1977, Thm. 1). On the days at Krems, the EM algorithm started from the k-means cluster centroids stops after 273 iterations with cluster probabilities 0.63 and 0.37 and means $(3.1, 12.4)$ and $(14.2, 27.5)$. For 58 of the 366 days, no degree of belonging reaches 0.8; 23 of them are in May, and none is in January, February, November, or December.

A soft clustering is turned into a hard clustering by assigning each data point to the cluster with the largest degree of belonging, and $1 - \max_{c=1,\ldots,k} u_c^{(r)}$ measures how far that assignment is from certain (Murphy, 2012, Sect. 11.2.3). For the days at Krems, this assignment agrees with k-means on 335 of the 366 days. Fig. 1(a) shows why the two disagree: the curve on which the two degrees of belonging (2) are equal is not the decision boundary of k-means, the perpendicular bisector of the two cluster centroids. It bends around the warm cluster, because the warm days scatter less than the cold ones and the covariance matrix of the warm cluster is smaller.

Hard clustering is a limit of the GMM. Fix every cluster probability at $1/k$ and every covariance matrix at $\sigma^2\mathbf{I}$, a multiple of the identity matrix with one

common σ^2; this shrinks the hypothesis space, which is model pruning, one of the three routes that regularization distinguishes. The degree of belonging (2) then depends on the feature vector only through its squared Euclidean distances to the means,

$$u_c^{(r)} = \frac{\exp\big(-\big\|\mathbf{x}^{(r)} - \boldsymbol{\mu}^{(c)}\big\|_2^2/(2\sigma^2)\big)}{\sum_{c'=1}^{k} \exp\big(-\big\|\mathbf{x}^{(r)} - \boldsymbol{\mu}^{(c')}\big\|_2^2/(2\sigma^2)\big)}, \tag{3}$$

the softmax function of the negative squared Euclidean distances divided by $2\sigma^2$. As σ^2 shrinks towards 0, the degree of belonging (3) of the nearest mean tends to 1 and all others to 0, so the vector (1) tends to the vertex of the nearest-cluster centroid assignment, and the EM algorithm becomes Lloyd's algorithm for k-means (Bishop, 2006, Sect. 9.3.2; Hastie et al., 2009, Sect. 14.3.7). With the means at the k-means cluster centroids of the days at Krems, 26 days have a degree of belonging between 0.2 and 0.8 for $\sigma^2 = 25$, three days for $\sigma^2 = 4$, and none for $\sigma^2 = 1$, while for every σ^2 the largest degree of belonging names the nearest cluster centroid. Fig. 1(b) draws the degree of belonging to the cold cluster along the line through the two cluster centroids, parametrized by s with $s = 0$ at the cold and $s = 1$ at the warm cluster centroid. Along this line, k-means jumps from 1 to 0 at $s = 1/2$, the GMM crosses 1/2 later and gradually, and the constrained GMM with $\sigma^2 = 4$ is close to the jump.

Fuzzy c-means delivers degrees of belonging without a probabilistic model (Bezdek, 1981). It minimizes a weighted clustering error,

$$\frac{1}{m}\sum_{r=1}^{m}\sum_{c=1}^{k}\big(u_c^{(r)}\big)^{\varphi}\Big\|\mathbf{x}^{(r)} - \mathbf{w}^{(c)}\Big\|_2^2, \tag{4}$$

jointly over the cluster centroids $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)} \in \mathbb{R}^d$ and the degrees of belonging, each vector (1) constrained to the probability simplex, with a fuzzifier $\varphi > 1$. For fixed cluster centroids, the minimizing degrees of belonging are proportional to $\big\|\mathbf{x}^{(r)} - \mathbf{w}^{(c)}\big\|_2^{-2/(\varphi-1)}$, normalized to sum to one; for fixed degrees of belonging, the minimizing cluster centroid $\mathbf{w}^{(c)}$ is the average of the feature vectors weighted by $(u_c^{(r)})^{\varphi}$. Alternating the two updates is a fixed-point iteration: each update minimizes the objective (4) exactly in one block of variables with the other fixed, so no iteration increases the objective, and a fixed point is a pair of cluster centroids and degrees of belonging at which neither update changes anything. As φ tends to 1, the degrees of belonging tend to the vertices of the simplex and the objective (4) to the clustering error of k-means. With $\varphi = 1.05$ on the days at Krems, the cluster centroids lie within 0.05 of those of k-means and the objective is 42.0, the clustering error of k-means. For 358 of the 366 days, the largest degree of belonging exceeds 0.99. With $\varphi = 2$, 66 days have no degree of belonging of 0.8 or more, and with $\varphi = 10$ every degree of belonging lies between 0.26 and 0.74.

The degrees of belonging of fuzzy c-means decay like the reciprocal of a squared Euclidean distance, so far away from every cluster centroid they approach $1/k$. At the ends of the line in Fig. 1(b), the degree of belonging to the

cold cluster is 0.91 at $s = -1/2$ and 0.09 at $s = 3/2$, where the GMM gives 1 and 0 to within 0.02.

In image segmentation by clustering the patches of a photograph by their colors, a patch on the horizon between a mountain and the sky is partly both, and its degree of belonging vector records both shares while a hard clustering must choose one. A soft clustering of the days at Krems records the transitional days of spring and autumn, which a hard clustering into a cold and a warm regime hides.

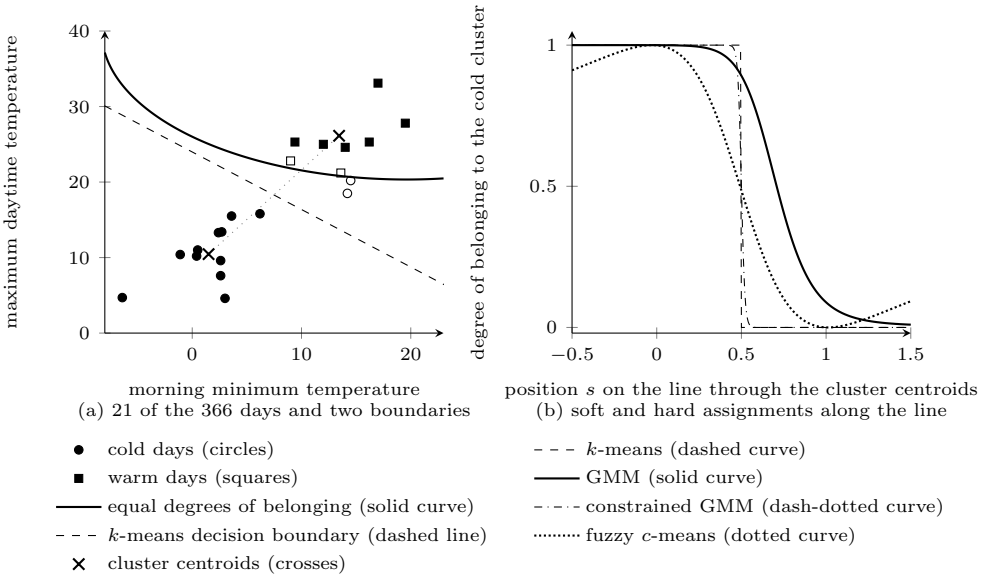


Figure 1: (a) Every 18th of the 366 days of 2024 at Krems, 21 days, drawn as cold days (circles) or warm days (squares) by the larger of their two degrees of belonging (2) under the GMM; the four open markers are days whose larger degree of belonging is below 0.8. The solid curve is the set of feature vectors with equal degrees of belonging to both clusters, the dashed line the decision boundary of k-means, the crosses its two cluster centroids, and the thin dotted segment between them the line of panel (b). (b) The degree of belonging to the cold cluster along the line through the two cluster centroids: k-means (dashed curve) jumps at $s = 1/2$, the GMM (solid curve) and fuzzy c-means with $\varphi = 2$ (dotted curve) change gradually, and the constrained GMM (3) with $\sigma^2 = 4$ (dash-dotted curve) is close to the jump. Data generated by `pythondemos/softclustering.py`

Cross References

clustering, cluster, degree of belonging, hard clustering, probability simplex, GMM, posterior distribution, EM, maximum likelihood, clustering error, k-

means, Lloyd's algorithm, cluster centroid, softmax function, fixed-point iteration, decision boundary, image segmentation, spectral clustering

References

1. Bezdek (1981). Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York, NY, USA.

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

3. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

4. 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.

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.