degree of belonging

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

membership degree, responsibility

Abstract

A degree of belonging is a number that indicates the extent to which a data point belongs to a cluster, one for each pair of a data point and a cluster of a clustering. In the most common formulation, the degrees of belonging of one data point are nonnegative and sum to one, so they form a vector of the probability simplex and a probability mass function (pmf) over the cluster indices. Hard clustering is the special case in which this vector has one entry equal to one and all others equal to zero, as delivered by k-means. A principled construction of degrees of belonging strictly between zero and one is the posterior distribution of the cluster index under a Gaussian mixture model (GMM), which the expectation–maximization (EM) algorithm computes in its E-step. The degrees of belonging in turn determine the clusters: the mean of a cluster is the average of the feature vectors weighted by them, and with degrees of belonging of zero or one this weighted average is the averaging step of Lloyd's algorithm. As the common variance of the clusters of a GMM shrinks to zero, the degrees of belonging turn into zero or one and soft clustering turns into hard clustering.

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 cold cluster and a warm cluster. A day in January is a cold day and a day in July a warm day, but 28 May 2024, with a morning minimum of 11.1 and a maximum of 21.3 degrees, lies between the two. A degree of belonging $u_c^{(r)}$ is a number that indicates the extent to which data point r belongs to cluster c, for $c = 1, \ldots, k$ and $r = 1, \ldots, m$. For 28 May 2024, the Gaussian mixture model (GMM) of Fig. 2 delivers the degrees of belonging 0.50 to the cold and 0.50 to the warm cluster, while a day in January obtains 1.00 and 0.00. In the most common formulation, the degrees of belonging of one data point are nonnegative

and sum to one (Bezdek, 1981; Ruspini, 1969),

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

The vector $\mathbf{u}^{(r)} := \left(u_1^{(r)}, \ldots, u_k^{(r)}\right)^\top$ then lies in the probability simplex Δ^{k-1}. It is the probability mass function (pmf) of a random variable (RV) with values in $\{1, \ldots, k\}$, the cluster index of the data point. Possibilistic clustering drops the sum constraint of the simplex condition (1) and lets each degree of belonging vary in $[0, 1]$ on its own (Krishnapuram and Keller, 1993). Fig. 1 sketches three data points and their degrees of belonging to two overlapping clusters.

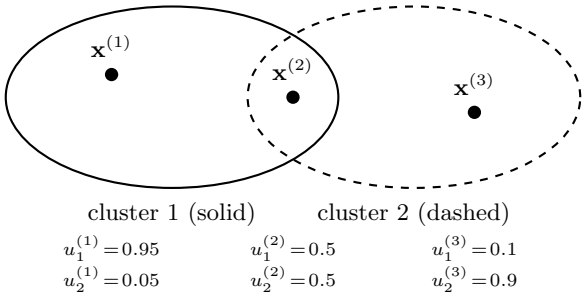


Figure 1: Three data points with different degrees of belonging to two overlapping clusters, cluster 1 (solid ellipse) and cluster 2 (dashed ellipse). The feature vector $\mathbf{x}^{(1)}$ lies deep inside cluster 1 and $\mathbf{x}^{(3)}$ deep inside cluster 2, while $\mathbf{x}^{(2)}$ lies in the overlap with equal degrees of belonging. In every case $u_1^{(r)} + u_2^{(r)} = 1$, the simplex condition (1)

Hard clustering is the special case in which every data point belongs to exactly one cluster. With $\widehat{c}^{(r)}$ the index of the cluster that data point r is assigned to, its degrees of belonging are

$$u_c^{(r)} = \begin{cases} 1 & \text{for } c = \widehat{c}^{(r)}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

so the vector $\mathbf{u}^{(r)}$ is a vertex of the probability simplex Δ^{k-1} (Bishop, 2006, Sect. 9.1). k-means delivers such a hard clustering: $\widehat{c}^{(r)}$ is the index of the cluster centroid nearest to $\mathbf{x}^{(r)}$ in Euclidean distance. For the days at Krems and two cluster centroids, k-means assigns 192 days to the cold cluster with cluster centroid $(1.5, 10.5)$ and 174 days to the warm cluster with cluster centroid $(13.4, 26.1)$. The dashed curve in Fig. 2(b) shows these hard assignments over the year: they jump between 0 and 1 from one spring day to the next.

Soft clustering methods assign a data point to several clusters with degrees of belonging strictly between 0 and 1. A principled construction uses a probabilistic model for the data points, the GMM. Under it, the cluster index

$I^{(r)} \in \{1, \ldots, k\}$ of data point r is an unobserved RV with $\mathbb{P}\left(I^{(r)} = c\right) = p_c$. Given $I^{(r)} = c$, the feature vector $\mathbf{x}^{(r)}$ is drawn from a multivariate normal distribution with mean $\boldsymbol{\mu}^{(c)}$ and covariance matrix $\mathbf{C}^{(c)}$. The degree of belonging is then the posterior distribution of the cluster index given the observed feature vector,

$$u_c^{(r)} := \mathbb{P}\left(I^{(r)} = c \mid \mathbf{x}^{(r)}\right) = \frac{p_c \, \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right)}{\sum_{c'=1}^{k} p_{c'} \, \mathcal{N}\left(\boldsymbol{\mu}^{(c')}, \mathbf{C}^{(c')}\right)\left(\mathbf{x}^{(r)}\right)}, \tag{3}$$

the responsibility of cluster c for the data point (Bishop, 2006, Sect. 9.2; Murphy, 2012, Sect. 11.2.3). The posterior degrees of belonging (3) are the probabilities of the k values of one RV, so they satisfy the simplex condition (1) by construction. The model parameters p_c, $\boldsymbol{\mu}^{(c)}$, and $\mathbf{C}^{(c)}$ of the GMM are learned by maximum likelihood. This is an instance of empirical risk minimization (ERM): the loss of a data point is the negative logarithm of the probability density function (pdf) of the GMM at its feature vector, and the clustering error is the average of this loss over the dataset. For the days at Krems, a two-component GMM learned by the expectation–maximization (EM) algorithm has the means $(3.1, 12.3)$ and $(14.2, 27.5)$ and the cluster probabilities 0.63 and 0.37. Under it, 58 of the 366 days have no degree of belonging above 0.8, 23 of them in May and none between November and February (Fig. 2).

The degrees of belonging determine the clusters in turn. In the M-step of the EM algorithm, the mean of cluster c is the average of the feature vectors weighted by the degrees of belonging. The cluster probability is the average of the degrees of belonging themselves. In formulas,

$$\boldsymbol{\mu}^{(c)} = \frac{\sum_{r=1}^{m} u_c^{(r)} \, \mathbf{x}^{(r)}}{\sum_{r=1}^{m} u_c^{(r)}} \qquad \text{and} \qquad p_c = \frac{1}{m} \sum_{r=1}^{m} u_c^{(r)} \quad \text{for } c = 1, \ldots, k. \tag{4}$$

The covariance matrix $\mathbf{C}^{(c)}$ is the covariance matrix of the feature vectors under the same weighting (Bishop, 2006, Sect. 9.2.2). A day with the degrees of belonging 0.50 and 0.50 counts half towards each mean. With the degrees of belonging (2) of a hard clustering, the weighted average (4) is the mean of the data points assigned to cluster c, which is the averaging step of Lloyd's algorithm for k-means, and p_c is the share of data points in that cluster. The EM algorithm alternates the posterior (3) with the weighted averages (4), so the degrees of belonging are recomputed in every iteration: they are not model parameters but a function of the model parameters and of the feature vector. At the fixed point of the EM algorithm on the days at Krems, each mean equals the average of the days weighted by their degrees of belonging and differs from the k-means cluster centroid of the same cluster by at most 1.9 degrees.

Hard clustering is also a limit of the posterior degrees of belonging (3). Fix every cluster probability at $1/k$ and every covariance matrix at $\sigma^2 \mathbf{I}$, a multiple of the identity matrix $\mathbf{I}$ with one common variance σ^2. The posterior (3) then depends on $\mathbf{x}^{(r)}$ only through its squared Euclidean distances to the means. As σ^2 shrinks towards 0, it concentrates on the nearest mean: every degree of

belonging tends to 0 or 1, and the EM algorithm reduces to Lloyd's algorithm (Bishop, 2006, Sect. 9.3.2; Hastie et al., 2009, Sect. 14.3.7). For the days at Krems, with the two k-means cluster centroids as means and $\sigma^2 = 1$, no day has both degrees of belonging above 0.2, and every degree of belonging is within 0.01 of its k-means value (2). With $\sigma^2 = 50$, 16 percent of the days have both degrees above 0.2. With $\sigma^2 = 1000$, every day has both above 0.2 and no degree of belonging exceeds 0.63.

The degrees of belonging of a data point are the value of a hypothesis at its feature vector. The posterior (3) is a map $h : \mathbb{R}^d \to \Delta^{k-1}$ determined by the model parameters of the GMM, and the assignment rule of k-means is a map from $\mathbb{R}^d$ to the vertices of Δ^{k-1} determined by the cluster centroids. Either map applies to a day that did not belong to the dataset the clusters were learned from, which is how a clustering generalizes. A hard clustering is derived from degrees of belonging by assigning each data point to the cluster with the largest one, and $1 - \max_{c=1,\ldots,k} u_c^{(r)}$ measures how uncertain that assignment is (Murphy, 2012, Sect. 11.2.3). For the days at Krems, the largest posterior degree of belonging names the k-means cluster on 91.5 percent of the days. In medical image segmentation, the degrees of belonging of a pixel to the tissue types let a pixel at the boundary between two tissues belong partly to both, where a hard clustering must pick one (Pham et al., 2000).

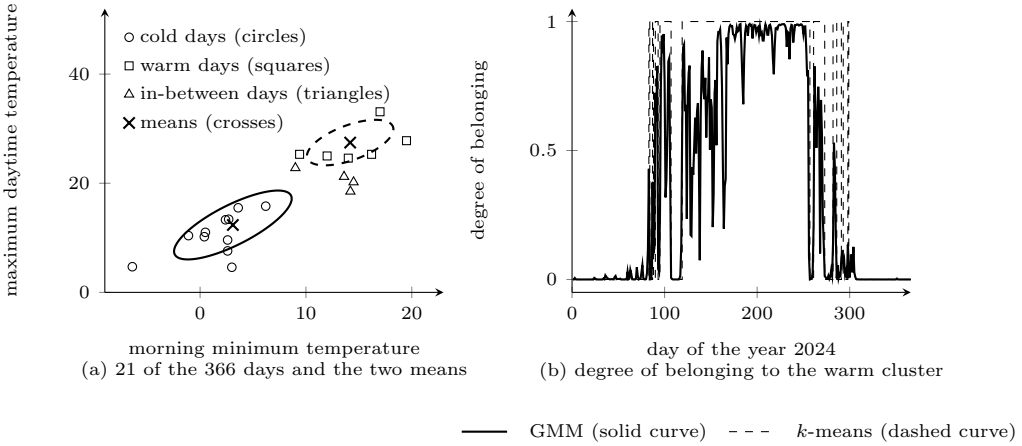


Figure 2: (a) Every 18th of the 366 days of 2024 at Krems, 21 days, drawn by their posterior degrees of belonging (3) under the two-component GMM: cold days (circles) and warm days (squares) have a degree of belonging above 0.8 to one cluster, in-between days (triangles) have both below 0.8. The crosses are the two means, and the ellipses are the level sets of the two multivariate normal distributions at one standard deviation: the cold one (solid ellipse) and the warm one (dashed ellipse). (b) The degree of belonging to the warm cluster of every day of the year: the posterior under the GMM (solid curve) and the 0 or 1 of k-means (dashed curve). Data generated by `pythondemos/dob.py`

Cross References

cluster, clustering, soft clustering, hard clustering, cluster centroid, GMM, EM, k-means, Lloyd's algorithm, posterior distribution, probability simplex, pmf, clustering error

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. Krishnapuram and Keller (1993). A possibilistic approach to clustering. IEEE Transactions on Fuzzy Systems, 1(2), pp. 98–110.

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

5. Pham et al. (2000). Current methods in medical image segmentation. Annual Review of Biomedical Engineering, 2, pp. 315–337.

6. Ruspini (1969). A new approach to clustering. Information and Control, 15(1), pp. 22–32.

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