Soft Clustering

Autor und Herausgeber

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

Zusammenfassung

Das Soft Clustering ordnet jeden Datenpunkt jedem Cluster mit einem Zugehörigkeitsgrad zu, einer Zahl zwischen null und eins, sodass die Zugehörigkeitsgrade eines Datenpunkts einen Vektor im Wahrscheinlichkeitssimplex bilden. Das hartes Clustering ist der Spezialfall, in dem dieser Vektor eine Ecke des Simplex ist, wobei ein Eintrag gleich eins ist und die anderen gleich null sind. Für numerische Merkmalsvektoren ist die prinzipiengeleitete Wahl der Zugehörigkeitsgrade die A-posteriori-Verteilung des Cluster-Index unter einem Gaußsches Mischmodell (GMM), dessen Modellparameter durch den EM-Algorithmus (EM)-Algorithmus gelernt werden. Bei gleichen Cluster-Wahrscheinlichkeiten und Kovarianzmatrizen $\sigma^2\mathbf{I}$ streben die Zugehörigkeitsgrade gegen die Ecken, wenn σ^2 schrumpft, und der EM-Algorithmus geht in den Lloyd-Algorithmus für k-Means über. Fuzzy-c-Means liefert Zugehörigkeitsgrade ohne ein Wahrscheinlichkeitsmodell, indem es einen gewichteten Clusteringfehler über dem Simplex minimiert, und reduziert sich ebenfalls auf k-Means, wenn sein Fuzzifier gegen eins strebt. Die Zuordnung jedes Datenpunkts zu dem Cluster mit dem größten Zugehörigkeitsgrad überführt ein Soft Clustering in ein hartes Clustering, und eins minus diesem größten Zugehörigkeitsgrad misst, wie weit die Zuordnung von der Gewissheit entfernt ist.

Definition

Betrachtet werden die 365 Tage des Jahres 2024 in Krems an der Donau (Österreich), jeweils ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$, der die Lufttemperatur um 06:00 Uhr und die Temperatur um 15:00 Uhr (UTC) des Tages enthält. K-Means mit $k = 2$ Cluster ordnet jeden Tag dem näheren von zwei Clusterzentren zu, $(3.9, 9.0)$ für 198 kalte Tage und $(17.5, 24.9)$ für 167 warme Tage. Ein Tag im April, dessen Temperaturen auf halbem Weg zwischen den beiden Clusterzentren liegen, wird ebenso zu einem kalten oder einem warmen Tag, und nichts hält fest, dass es sich um einen Grenzfall handelte. Das Soft Clustering bewahrt diese Information: Es ordnet jeden Datenpunkt jedem Cluster mit einem Zugehörigkeitsgrad $u_c^{(r)}$ zu, einer Zahl zwischen 0 und 1 für den Datenpunkt $r = 1, \ldots, m$ und das Cluster $c = 1, \ldots, k$ (Bishop, 2006, Abschn. 9.3.2; Murphy, 2012, Abschn. 11.2.3). Die Zugehörigkeitsgrade eines

Datenpunkts bilden den Vektor

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

einen Punkt des Wahrscheinlichkeitssimplex Δ^{k-1}. Ein hartes Clustering ist der Spezialfall, in dem jeder Vektor (1) eine Ecke des Simplex ist: Ein Eintrag ist 1 und die anderen $k-1$ Einträge sind 0. Ein Verfahren für das Soft Clustering lernt eine Hypothese $h : \mathcal{X} \to \Delta^{k-1}$, die einen Merkmalsvektor auf einen Zugehörigkeitsgrad-Vektor abbildet, sodass sie für einen Tag ausgewertet werden kann, der nicht zum Datensatz gehörte. Ein Verfahren für das hartes Clustering lernt eine Abbildung ausschließlich in die k Ecken.

Für Datenpunkte mit numerischen Merkmalsvektoren ist die prinzipiengeleitete Wahl der Zugehörigkeitsgrade ein Wahrscheinlichkeitsmodell, das Gaußsches Mischmodell (GMM). Es erzeugt einen Datenpunkt, indem es einen Cluster-Index I mit den Wahrscheinlichkeiten $p_1, \ldots, p_k$ zieht und anschließend den Merkmalsvektor aus der multivariate Normalverteilung dieses Clusters mit dem Mittelwert $\boldsymbol{\mu}^{(c)}$ und der Kovarianzmatrix $\mathbf{C}^{(c)}$. Der Zugehörigkeitsgrad des Datenpunkts $\mathbf{x}^{(r)}$ zum Cluster c ist die A-posteriori-Verteilung seines Cluster-Index,

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

was Bishop (2006), Abschn. 9.2 als die Zuständigkeit (engl. *responsibility*) des Clusters c für den Datenpunkt bezeichnet. Die Modellparameter werden mittels der maximale Likelihood durch den EM-Algorithmus (EM)-Algorithmus gelernt, dessen E-Schritt die Zugehörigkeitsgrade (2) berechnet und dessen M-Schritt die Modellparameter als mit diesen gewichtete Mittelwerte neu berechnet (Bishop, 2006, Abschn. 9.2.2). Die maximale Likelihood ist empirische Risikominimierung (ERM) mit dem negativen Logarithmus der Wahrscheinlichkeitsdichte als Verlust, sodass die gelernten Modellparameter einen Clusteringfehler minimieren, und der EM-Algorithmus ist eine Fixpunktiteration, die diesen niemals vergrößert (Dempster et al., 1977, Thm. 1). Auf den Tagen in Krems stoppt der an den k-Means-Clusterzentren gestartete EM-Algorithmus nach 315 Iterationen mit den Cluster-Wahrscheinlichkeiten 0.69 und 0.31 sowie den Mittelwerte $(6.2, 11.6)$ und $(19.0, 27.0)$. Für 60 der 365 Tage erreicht kein Zugehörigkeitsgrad den Wert 0.8; 22 davon liegen im Mai, und keiner liegt im Januar, Februar, November oder Dezember.

Ein Soft Clustering wird in ein hartes Clustering überführt, indem jeder Datenpunkt dem Cluster mit dem größten Zugehörigkeitsgrad zugeordnet wird, und $1 - \max_{c=1,\ldots,k} u_c^{(r)}$ misst, wie weit diese Zuordnung von der Gewissheit entfernt ist (Murphy, 2012, Abschn. 11.2.3). Für die Tage in Krems stimmt diese Zuordnung an 316 der 365 Tage mit k-Means überein. Abb. 1(a) zeigt, warum die beiden voneinander abweichen: Die Kurve, auf der die beiden Zugehörigkeitsgrade (2) gleich sind, ist nicht die Entscheidungsgrenze von k-Means,

die Mittelsenkrechte der beiden Clusterzentren. Sie krümmt sich um das warme Cluster, da die warmen Tage weniger streuen als die kalten und die Kovarianzmatrix des warmen Clusters kleiner ist.

Das hartes Clustering ist ein Grenzfall des GMM. Legt man jede Cluster-Wahrscheinlichkeit auf $1/k$ und jede Kovarianzmatrix auf $\sigma^2\mathbf{I}$ fest, ein Vielfaches der Einheits-Matrix mit einem gemeinsamen σ^2, so verkleinert dies den Hypothesenraum, was dem Modellpruning entspricht, einem der drei Wege, welche die Regularisierung unterscheidet. Der Zugehörigkeitsgrad (2) hängt dann vom Merkmalsvektor nur über dessen quadrierte euklidische Abstände zu den Mittelwerte ab,

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

der Softmax-Funktion der negativen quadrierten euklidische Abstände geteilt durch $2\sigma^2$. Wenn σ^2 gegen 0 schrumpft, strebt der Zugehörigkeitsgrad (3) des nächstgelegenen Mittelwerts gegen 1 und alle anderen streben gegen 0, sodass der Vektor (1) gegen die Ecke der Zuordnung zum nächstgelegenen Clusterzentrum strebt und der EM-Algorithmus in den Lloyd-Algorithmus für k-Means übergeht (Bishop, 2006, Abschn. 9.3.2; Hastie et al., 2009, Abschn. 14.3.7). Liegen die Mittelwerte bei den k-Means-Clusterzentren der Tage in Krems, so weisen für $\sigma^2 = 25$ genau 4 Tage einen Zugehörigkeitsgrad zwischen 0.2 und 0.8 auf, drei Tage für $\sigma^2 = 4$ und keiner für $\sigma^2 = 1$, während für jedes σ^2 der größte Zugehörigkeitsgrad den nächstgelegenen Clusterzentrum benennt. Abb. 1(b) stellt den Zugehörigkeitsgrad zum kalten Cluster entlang der Geraden durch die beiden Clusterzentren dar, parametrisiert durch s mit $s = 0$ am kalten und $s = 1$ am warmen Clusterzentrum. Entlang dieser Geraden springt k-Means bei $s = 1/2$ von 1 auf 0, das GMM überquert $1/2$ später und allmählich, und das eingeschränkte GMM mit $\sigma^2 = 4$ liegt nahe am Sprung.

Fuzzy-c-Means liefert Zugehörigkeitsgrade ohne ein Wahrscheinlichkeitsmodell (Bezdek, 1981). Es minimiert einen gewichteten Clusteringfehler,

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

gemeinsam über die Clusterzentren $\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)} \in \mathbb{R}^d$ und die Zugehörigkeitsgrade, wobei jeder Vektor (1) auf das Wahrscheinlichkeitssimplex beschränkt ist, mit einem Fuzzifier $\varphi > 1$. Für festgehaltene Clusterzentren sind die minimierenden Zugehörigkeitsgrade proportional zu $\big\|\mathbf{x}^{(r)} - \mathbf{w}^{(c)}\big\|_2^{-2/(\varphi-1)}$, normiert auf die Summe eins; für festgehaltene Zugehörigkeitsgrade ist der minimierende Clusterzentrum $\mathbf{w}^{(c)}$ das mit $(u_c^{(r)})^{\varphi}$ gewichtete Mittel der Merkmalsvektoren. Das Abwechseln der beiden Aktualisierungen ist eine Fixpunktiteration: Jede Aktualisierung minimiert die Zielfunktion (4) exakt in einem Variablenblock bei festgehaltenem anderem Block, sodass keine Iteration die Zielfunktion vergrößert, und ein Fixpunkt ist ein Paar aus Clusterzentren und Zugehörigkeitsgrade, bei dem keine der beiden Aktualisierungen etwas ändert. Wenn φ gegen 1 strebt,

streben die Zugehörigkeitsgrade gegen die Ecken des Simplex und die Zielfunktion (4) strebt gegen den Clusteringfehler von k-Means. Mit $\varphi = 1.05$ auf den Tagen in Krems liegen die Clusterzentren innerhalb von 0.05 bei jenen von k-Means und die Zielfunktion beträgt 44.8, was dem Clusteringfehler von k-Means entspricht. Für 356 der 365 Tage übersteigt der größte Zugehörigkeitsgrad den Wert 0.99. Mit $\varphi = 2$ weisen 63 Tage keinen Zugehörigkeitsgrad von 0.8 oder mehr auf, und mit $\varphi = 10$ liegt jeder Zugehörigkeitsgrad zwischen 0.27 und 0.73.

Die Zugehörigkeitsgrade von Fuzzy-c-Means fallen wie der Kehrwert einer quadrierten euklidischer Abstand ab, sodass sie sich weit entfernt von jedem Clusterzentrum dem Wert $1/k$ annähern. An den Enden der Geraden in Abb. 1(b) beträgt der Zugehörigkeitsgrad zum kalten Cluster 0.91 bei $s = -1/2$ und 0.09 bei $s = 3/2$, wo das GMM bis auf 0.02 genau die Werte 1 und 0 liefert.

Bei der Bildsegmentierung durch das Clustering der Ausschnitte eines Fotos nach ihren Farben ist ein Ausschnitt am Horizont zwischen einem Berg und dem Himmel teilweise beides, und sein Zugehörigkeitsgrad-Vektor erfasst beide Anteile, während ein hartes Clustering sich für einen entscheiden muss. Ein Soft Clustering der Tage in Krems erfasst die Übergangstage im Frühling und Herbst, die ein hartes Clustering in ein kaltes und ein warmes Regime verbirgt.

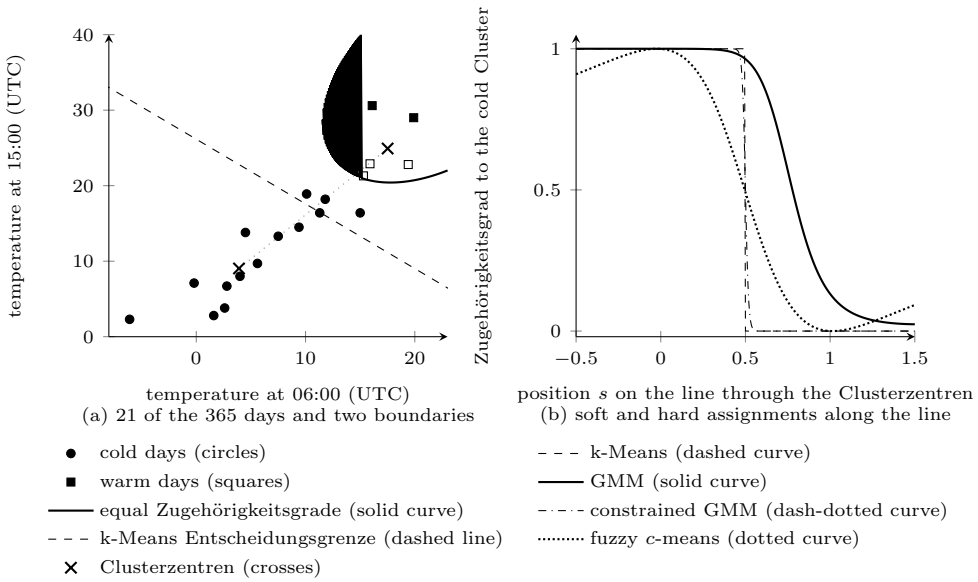


Abbildung 1: (a) Jeder 18 der 365 Tage des Jahres 2024 in Krems, 21 Tage, eingezeichnet als kalte Tage (Kreise) oder warme Tage (Quadrate) anhand des größeren ihrer beiden Zugehörigkeitsgrade (2) unter dem GMM; die vier ungefüllten Symbole sind Tage, deren größerer Zugehörigkeitsgrad unter 0.8 liegt. Die durchgezogene Kurve ist die Menge der Merkmalsvektoren mit gleichen Zugehörigkeitsgrade zu beiden Clustern, die gestrichelte Gerade die Entscheidungsgrenze von k-Means, die Kreuze dessen beide Clusterzentren und das dünne gepunktete Segment dazwischen die Gerade aus Teilbild (b). (b) Der Zugehörigkeitsgrad zum kalten Cluster entlang der Geraden durch die beiden Clusterzentren: k-Means (gestrichelte Kurve) springt bei $s = 1/2$, das GMM (durchgezogene Kurve) und Fuzzy-c-Means mit $\varphi = 2$ (gepunktete Kurve) ändern sich allmählich, und das eingeschränkte GMM (3) mit $\sigma^2 = 4$ (strichpunktierte Kurve) liegt nahe am Sprung. Daten erzeugt durch `pythondemos/softclustering.py`

Synonyme: Fuzzy-Clustering, weiches Clustering.

Siehe auch: Clustering, Cluster, Zugehörigkeitsgrad, hartes Clustering, Wahrscheinlichkeitssimplex, GMM, A-posteriori-Verteilung, EM, maximale Likelihood, Clusteringfehler, k-Means, Lloyd-Algorithmus, Clusterzentrum, Softmax-Funktion, Fixpunktiteration, Entscheidungsgrenze, Bildsegmentierung, Spektrales Clustering.

Querverweise

Literatur

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.