# Clusterzentrum

## 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

Ein Clusterzentrum ist ein Vektor, das ein Cluster von Datenpunkteen repräsentiert: Unter allen Vektorenen minimiert es den summierten Verlust über die Datenpunktee des Clusters, wobei der Verlust misst, wie schlecht ein Vektor einen Datenpunkt ersetzt. Unter der quadrierten euklidischer Abstand ist das Clusterzentrum der Mittelwert der Merkmalsvektorenen im Cluster, unter der euklidischer Abstand deren geometric median (GM), und unter der Summe der absoluten Differenzen der median jedes Merkmals. K-Means weist einen Datenpunkt seinem nächsten Clusterzentrum zu, eine Regel, die eine linearer Klassifikator ist, deren Entscheidungsgrenze zwischen zwei Clusterzentren die Hyperebene durch ihren Mittelpunkt und senkrecht zu ihrer Differenz ist. Der Mittelungsschritt des Lloyd's algorithm berechnet jedes Clusterzentrum als den Mittelwert seines Clusters neu, und an einem Fixpunkt ist jedes Clusterzentrum der Mittelwert der ihm nächsten Datenpunktee. Der Mittelwert bewegt sich mit jedem Datenpunkt, sodass ein Ausreißer ihn um die Verschiebung geteilt durch die Größe des Clusters verschiebt, während sich der GM und der koordinatenweise median wenig bewegen. Eine soft clustering-Methode, wie ein Gaußsches Mischmodell (GMM), verwendet den mit den Mittelwerteen gewichteten Zugehörigkeitsgrad als ihr Clusterzentrum. Das Ersetzen jedes Datenpunkts durch sein Clusterzentrum komprimiert einen Datensatz auf $k$ Vektoren und einen Cluster-Index pro Datenpunkt.

## Definition

Man betrachte die 365 Tage des Jahres 2024 in Krems an der Donau (Österreich), jeder ein Datenpunkt mit einem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^2$, der den Messwert um 06:00 Uhr und die Temperatur um 15:00 Uhr (UTC) des Tages enthält. K-Means mit $k = 2$ zerlegt diese Tage in 197 kalte Tage und 168 warme Tage, und es beschreibt jedes Cluster durch einen einzigen Vektor: einen typischen kalten Tag, $(3.9, 9.0)$, und einen typischen warmen Tag, $(17.5, 24.9)$. Ein solcher Vektor ist ein Clusterzentrum. Eine hartes Clustering-Methode, die jedes Cluster $c = 1, \ldots, k$ durch ein Clusterzentrum $\mathbf{w}^{(c)} \in \mathbb{R}^d$ repräsentiert, weist einen Datenpunkt dem Cluster zu, dessen Clusterzentrum am nächsten liegt (Bishop, 2006, Sect. 9.1; Hastie et al., 2009, Sect. 14.3.6), wie Abb. 1(a) für die Tage in Krems zeigt.

Welcher Vektor ein Cluster am besten repräsentiert, wird durch eine Verlustfunktion $L(\mathbf{x}, \mathbf{w})$ entschieden, die misst, wie schlecht ein Vektor $\mathbf{w}$ den Merkmalsvektor $\mathbf{x}$ eines Datenpunkts ersetzt. Das Clusterzentrum eines Clusters mit Indexmenge $\mathcal{C}^{(c)} \subseteq \{1, \ldots, m\}$ ist ein Vektor, der den summierten Verlust über seine Datenpunktee minimiert,

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

Für die quadrierte euklidischer Abstand $L(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|_2^2$ ist der summierte Verlust eine strictly convex quadratic function von $\mathbf{w}$, und sein eindeutiger Minimierer ist der Mittelwert der Merkmalsvektorenen,

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

da der Gradient des summierten Verlusts genau dort verschwindet, wo die Abweichungen $\mathbf{x}^{(r)} - \mathbf{w}$ sich zu Null summieren (MacQueen, 1967). Die beiden Clusterzentren der Tage in Krems sind die Mittelwertee (2) der kalten und der warmen Tage. Unter der euklidischer Abstand selbst, $L(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|_2$, ist der Minimierer des summierten Verlusts (1) der geometric median (GM) der Merkmalsvektorenen (Boyd and Vandenberghe, 2004, Sect. 8.7.1). Er hat keine geschlossene Form und wird durch Weiszfelds fixed-point iteration (Weiszfeld, 1937) berechnet,

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

ein Mittelwert der Merkmalsvektorenen, gewichtet mit dem Kehrwert ihrer Abstand zum aktuellen Iterationswert. Ein Fixpunkt der Iteration (3), der mit keinem Merkmalsvektor übereinstimmt, ist ein Punkt, an dem sich die EinheitsVektorenen zu den Merkmalsvektorenen zu Null summieren, was die Optimalitätsbedingung des GMs ist, und der summierte Verlust von einem Iterationswert zum nächsten nie zunimmt (Weiszfeld, 1937). Unter der Summe der absoluten Differenzen, $L(\mathbf{x}, \mathbf{w}) = \|\mathbf{x} - \mathbf{w}\|_1$, zerfällt der summierte Verlust in ein Problem pro Merkmal, sodass das Clusterzentrum den median jedes Merkmals über das Cluster enthält. Für die warmen Tage in Krems ergeben die drei Optionen $(17.5, 24.9)$, $(17.4, 24.8)$ und $(17.8, 24.9)$, die innerhalb von 0.4 Grad voneinander liegen; Weiszfelds Iteration, gestartet vom Mittelwert, stoppt nach 605 Iterationenen.

Für ein festes Cluster ist die Minimierung (1) empirische Risikominimierung (ERM) über den Hypothesenraum konstanter Abbildungen $h(\mathbf{x}) = \mathbf{w}$. Das Clusterzentrum ist die Hypothese, die einen Vektor für jeden Datenpunkt des Clusters vorhersagt, und der summierte Verlust ist sein Trainingsfehler bis auf den Faktor $1/|\mathcal{C}^{(c)}|$. Die $k$ Clusterzentren zusammen sind die Modellparameter der Hypothese $h(\mathbf{x}) = \mathbf{w}^{(c^\star(\mathbf{x}))}$, die von k-Means gelernt wird, die einen

Merkmalsvektor zu seinem nächsten Clusterzentrum sendet,

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

Der Durchschnitt des Verlusts $\|\mathbf{x} - h(\mathbf{x})\|_2^2$ über den Datensatz ist der clustering error. Die Nächste-Zentrum-Regel (4) ist eine linearer Klassifikator. Für zwei Clusterzentren ist ein Merkmalsvektor näher an $\mathbf{w}^{(1)}$ als an $\mathbf{w}^{(2)}$ genau dann, wenn

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

was sich aus der Expansion der beiden quadrierten euklidische Abständeen ergibt. Die Entscheidungsgrenze der linearen Regel (5) ist die Hyperebene durch den Mittelpunkt der beiden Clusterzentren und senkrecht zu ihrer Differenz (Bishop, 2006, Sect. 9.1), die gestrichelte Linie in Abb. 1(a). Jedes Cluster ist ein Schnittpunkt von Halbräumeen und somit eine konvexe Region. Die Regel (4) weist auch einen Datenpunkt außerhalb des Datensatzes zu. Ein Tag mit einem Messwert von 5 Grad um 06:00 Uhr morgens und 18 Grad um 15:00 Uhr liegt in Abständeen von 9.1 und 14.2 zu den beiden Clusterzentren der Tage in Krems, daher ist es ein kalter Tag. Dieselbe Regel dient der Klassifikation, wenn jede Klasse eines gelabelten Datensatzes durch ein oder mehrere Clusterzentren repräsentiert wird und ein neuer Datenpunkt das Label des nächsten erhält (Hastie et al., 2009, Sect. 13.2.1).

Lloyd's algorithm für k-Means wechselt die Zuweisung (4) jedes Datenpunkts mit einem Mittelungsschritt ab, der jedes Clusterzentrum als den Mittelwert (2) der ihm zugewiesenen Datenpunktee neu berechnet (Lloyd, 1982). An einem Fixpunkt ist jedes Clusterzentrum der Mittelwert der ihm nächsten Datenpunktee. Zwei Bedingungen gelten dann gleichzeitig: jedes Clusterzentrum minimiert den summierten Verlust (1) über sein Cluster, und jedes Cluster besteht aus den Datenpunkteen, die seinem Clusterzentrum am nächsten liegen (Hastie et al., 2009, Sect. 14.3.6). Für die Tage in Krems, gestartet von den Mittelwerteen der kälteren und der wärmeren Hälfte der Tage, erreicht Lloyd's algorithm diesen Fixpunkt nach 2 Iterationenen. Das Ändern der Verlustfunktion ändert nur den Mittelungsschritt: der GM ergibt $k$-Mediane, und die Beschränkung des Clusterzentrums auf einen der Datenpunktee seines Clusters ergibt $k$-Medoide, was nur die paarweisen Abständeen zwischen Datenpunkteen benötigt (Hastie et al., 2009, Sect. 14.3.10; Park and Jun, 2009; Shalev-Shwartz and Ben-David, 2014, Sect. 22.2). Für eine Bregman divergence als Verlustfunktion bleibt der Mittelwert (2) das Clusterzentrum (Banerjee et al., 2005, Prop. 1).
Der Mittelwert (2) bewegt sich mit jedem Datenpunkt seines Cluster: Das Hinzufügen eines Vektor zu einem Merkmalsvektor fügt denselben Vektor, geteilt durch $|\mathcal{C}^{(c)}|$, zum Mittelwert hinzu, unabhängig von den anderen Datenpunkte. Würde der 15:00 (UTC)-Messwert des wärmsten warmen Tages in Krems als 999.9 aufgezeichnet, ein Wert, den einige Archive zur Kennzeichnung fehlender Aufzeichnungen verwenden, würde sich der Mittelwert der 168 warmen Tage um $965.6/168 = 5.7$ Grad verschieben (Abb. 1(b)). Der GM und der koordina-

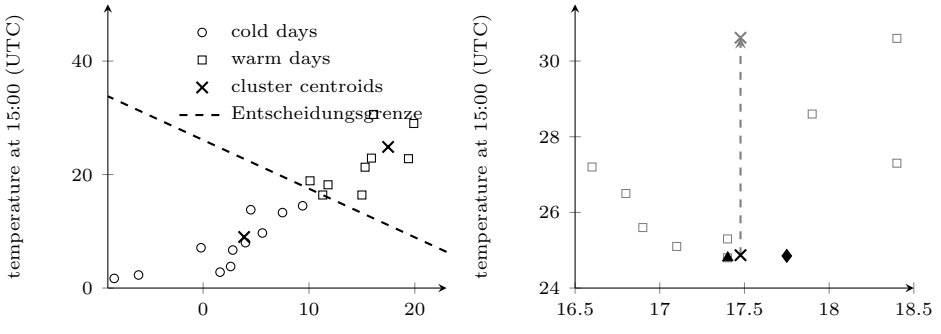
tenweise median würden sich jeweils um weniger als ein Hundertstel Grad verschieben. Die quadrierte euklidischer Abstand gewichtet die größten Abstände am stärksten, weshalb Clusterzentren, die aus einer unquadrierten Norm gebildet werden, weniger empfindlich gegenüber Ausreißer sind (Hastie et al., 2009, Sect. 14.3.10).

Eine soft clustering-Methode ordnet jeden Datenpunkt $r$ jedem Cluster $c$ mit einem Zugehörigkeitsgrad $u_c^{(r)} \in [0, 1]$ zu, wobei $\sum_{c=1}^{k} u_c^{(r)} = 1$ gilt. Sein Clusterzentrum ist der Mittelwert aller Merkmalsvektoren, gewichtet mit ihren Zugehörigkeitsgrade,

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

dessen Nenner die effektive Anzahl von Datenpunkte im Cluster ist. Für ein Gaußsches Mischmodell (GMM), das das Cluster $c$ durch eine multivariate Normalverteilung mit Mittelwert $\boldsymbol{\mu}^{(c)}$ und Kovarianzmatrix $\mathbf{C}^{(c)}$ darstellt, ist der Zugehörigkeitsgrad die A-posteriori-Verteilung des Cluster-Index, und der gewichtete Mittelwert (6) ist der M-Schritt des EM-Algorithmus (EM)-Algorithmus für $\boldsymbol{\mu}^{(c)}$ (Bishop, 2006, Sect. 9.2.2). Hartes Clustering ist der Fall $u_c^{(r)} \in \{0, 1\}$, bei dem der gewichtete Mittelwert (6) der Mittelwert (2) ist. Für die Tage in Krems stoppt der EM-Algorithmus, der von den beiden k-Means-Clusterzentren ausging, bei den Mittelwerte $(6.2, 11.6)$ und $(19.0, 27.0)$, die 3.5 bzw. 2.6 Grad von ihnen entfernt liegen. Von den 365 Tagen haben 60 keinen Zugehörigkeitsgrad über 0.8, und jeder von ihnen wird teilweise von beiden Clusterzentren gezählt.

Das Ersetzen jedes Datenpunkt durch sein Clusterzentrum komprimiert einen Datensatz auf $k$ Vektoren und einen Cluster-Index pro Datenpunkt, was Vektorquantisierung ist. Die 365 Tage in Krems werden zu zwei typischen Tagen und einem Bit pro Tag, und ein Foto, dessen Pixel nach Farbe geclustert sind, wird zu einer Palette von $k$ Farben und einem Palettenindex pro Pixel (Bishop, 2006, Sect. 9.1.1; Shalev-Shwartz and Ben-David, 2014, Sect. 22.2). Der clustering error ist die durchschnittliche Verzerrung dieser Kompression, und dessen Minimierung über die Clusterzentren ist die Aufgabe, aus der der Lloyd's algorithm entstand (Lloyd, 1982).




(a) 21 of the 365 days, two cluster centroids

(b) the cluster centroid of the warm days

Abbildung 1: (a) Jeder 18 der 365 Tage des Jahres 2024 in Krems, 21 Tage, dargestellt als kalte Tage (Kreise) oder warme Tage (Quadrate) durch das nähere der beiden Clusterzentren (Kreuze), die Mittelwerte der kalten und der warmen Tage; die gestrichelte Linie ist die Entscheidungsgrenze der Nächste-Zentroid-Regel, die Hyperebene durch den Mittelpunkt der beiden Clusterzentren und senkrecht zu deren Differenz. (b) Die 9 der 168 warmen Tage, die in das Fenster fallen (graue Quadrate), und das Clusterzentrum aller 168 warmen Tage unter drei Verlustfunktionen: ihr Mittelwert (schwarzes Kreuz) unter der quadrierten euklidischer Abstand, ihr GM (gefülltes Dreieck) unter der euklidischer Abstand und ihr koordinatenweiser median (gefüllte Raute) unter der Summe der absoluten Differenzen; der gestrichelte Pfeil führt vom Mittelwert zum Mittelwert, nachdem der 15:00 (UTC)-Messwert eines warmen Tages durch 999.9 ersetzt wurde (graues Kreuz), eine Verschiebung um 5.7 Grad, während sich die beiden medians jeweils um weniger als ein Hundertstel Grad verschieben. Daten generiert von `pythondemos/clustercentroid.py`

Siehe auch: Clustering, Cluster, clustering error, k-Means, Lloyd's algorithm, $k$-means++, hartes Clustering, soft clustering, Zugehörigkeitsgrad, GMM, EM, Mittelwert, median, GM, euklidischer Abstand, Norm, Ausreißer, linearer Klassifikator, Entscheidungsgrenze, Hyperebene, fixed-point iteration, ERM, Hypothese, Modellparameter, Bregman divergence, Merkmalsvektor.

## Querverweise

## Literatur

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.