Clusteringfehler

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

Der Clusteringfehler eines Clustering ist der Durchschnitt, über die Datenpunkte eines Datensatz, eines Verlust, der misst, wie schlecht ein Datenpunkt durch den ihm zugewiesenen Cluster repräsentiert wird. Für hartes Clustering-Methoden, die einen Cluster durch einen Clusterzentrum repräsentieren, wie zum Beispiel k-Means, ist der Verlust die quadrierte euklidischer Abstand zum nächsten Clusterzentrum. Die euklidischer Abstand selbst oder die Summe der absoluten Differenzen ergibt eine Variante, deren beste Clusterzentren Mediane anstelle von Mittelwerte sind und die weniger empfindlich gegenüber Ausreißer ist. Soft Clustering mit einem Gaußsches Mischmodell (GMM) verwendet den negativen Logarithmus der Wahrscheinlichkeitsdichte als Verlust, sodass sein Clusteringfehler die negative Log-Likelihood-Funktion pro Datenpunkt ist. In jedem Fall ist der Clusteringfehler der Trainingsfehler einer Hypothese, und seine Minimierung ist empirische Risikominimierung (ERM). Der Lloyd-Algorithmus und der EM-Algorithmus (EM)-Algorithmus erhöhen ihn nie, können aber an einem lokalen Minimum anhalten. Der Clusteringfehler nimmt mit der Anzahl der Cluster nicht zu und verschwindet, wenn jeder Datenpunkt sein eigener Cluster ist, sodass er diese Anzahl nicht wählen kann.

Definition

Man betrachte die 365 Tage eines Jahres 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. Eine Clustering-Methode zerlegt diese Tage in einige wenige Cluster, und Clustering-Methoden unterscheiden sich darin, wie sie einen Cluster repräsentieren. Eine Repräsentation ist ein Clusterzentrum, ein Vektor $\mathbf{w}^{(c)} \in \mathbb{R}^d$ für jeden Cluster $c = 1, \ldots, k$; jeder Datenpunkt wird dem Cluster mit dem nächsten Clusterzentrum zugewiesen, was das von k-Means gelieferte hartes Clustering ist. Ein natürliches Maß dafür, wie gut die Clusterzentren das Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ repräsentieren, ist der Durchschnitt über die Datenpunkte der minimalen quadrierten euklidischer Abstand zu den Clusterzentren (Bishop, 2006, Sect. 9.1, Eq. (9.1)),

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

Der Clusteringfehler von k-Means ist der Durchschnitt (1). Für zwei Clusterzentren und die Tage in Krems beträgt er 44.8; Abb. 1(a) zeigt die Abstände, die er für einige der Tage mittelt.

Letztendlich ist der Clusteringfehler (1) der Durchschnitt eines Verlust für eine spezielle Wahl der Verlustfunktion. Die von k-Means gelernte Hypothese ist die Abbildung $h(\mathbf{x}) = \mathbf{w}^{(c^\star(\mathbf{x}))}$, die einen Merkmalsvektor zu seinem nächsten Clusterzentrum sendet, wobei $c^\star(\mathbf{x}) = \arg\min_{c=1,\dots,k} \left\|\mathbf{x} - \mathbf{w}^{(c)}\right\|_2$ ist. Der Verlust eines Datenpunkt ist (Shalev-Shwartz and Ben-David, 2014, Sect. 22.2)

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

Ein Datenpunkt trägt hier kein Label: Der Verlust (2) vergleicht den Merkmalsvektor mit dem Clusterzentrum, der ihn repräsentiert. Der Clusteringfehler ist der Durchschnitt des Verlust (2) über das Datensatz, somit ist er der Trainingsfehler der Hypothese h, und die Wahl von Clusterzentren, die ihn minimieren, ist empirische Risikominimierung (ERM) über den Hypothesenraum aller Abbildungen dieser Form.

Die quadrierte euklidischer Abstand im Verlust (2) ist eine von mehreren Möglichkeiten. Das Ersetzen der quadrierten Norm durch die Norm selbst, $L\left(\mathbf{x}, h\right) = \left\|\mathbf{x} - h(\mathbf{x})\right\|_2$, oder durch die ℓ_1-Norm $\left\|\mathbf{x} - h(\mathbf{x})\right\|_1$, die Summe der absoluten Differenzen der Merkmale, behält die Zuweisung eines Datenpunkt zu seinem nächsten Clusterzentrum bei. Es ändert, welcher Clusterzentrum für einen gegebenen Cluster am besten ist. Unter der quadrierten euklidischer Abstand ist es der Mittelwert der Datenpunkte im Cluster, unter der euklidischer Abstand deren geometrischer Median (GM), berechnet durch Weiszfelds Fixpunktiteration (Weiszfeld, 1937), und unter der ℓ_1-Norm der Median jedes Merkmal separat. Ein einzelner weit entfernter Datenpunkt, ein Ausreißer, verschiebt den Mittelwert seines Cluster proportional zu seiner Abstand, verschiebt aber jeden Median nur wenig, sodass die Clusteringfehler, die aus einer unquadrierten Norm gebildet werden, weniger empfindlich gegenüber Ausreißer sind (Hastie et al., 2009, Sect. 14.3.10). Für die Tage in Krems und zwei Clusterzentren weisen die drei Optionen mehr als 98 Prozent der Tage demselben Cluster zu. Ihre Clusteringfehler, 44.8, 5.9 und 7.8, sind nicht vergleichbar, da der erste in Quadratgraden und die anderen beiden in Grad gemessen werden. Die quadrierte euklidischer Abstand kann auch durch eine Bregman-Divergenz ersetzt werden, für die der Mittelwert der beste Clusterzentrum bleibt (Banerjee et al., 2005).

Eine andere Wahl wird von Soft Clustering-Methoden verwendet, die einen Cluster durch eine Wahrscheinlichkeitsverteilung anstatt durch einen einzelnen Vektor repräsentieren und jeden Datenpunkt jedem Cluster mit einem Zugehörigkeitsgrad zuweisen. Ein Gaußsches Mischmodell (GMM) repräsentiert ein Clustering durch seine Modellparameter: Der Cluster c ist eine multivariate Normalverteilung mit Mittelwert $\boldsymbol{\mu}^{(c)}$ und Kovarianzmatrix $\mathbf{C}^{(c)}$, gewählt mit Wahrscheinlichkeit p_c. Der Clusteringfehler dieser Modellparameter ist die negative Log-Likelihood-Funktion des Datensatz pro Datenpunkt (Bishop, 2006,

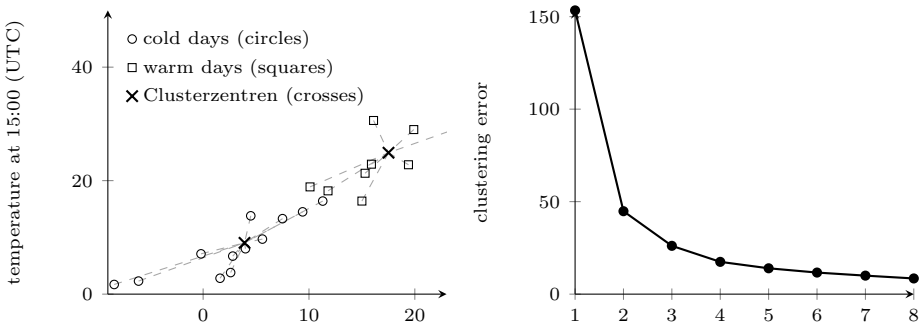
Sect. 9.2, Eq. (9.14)),

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

wieder der Durchschnitt eines Verlust, der negative Logarithmus der Wahrscheinlichkeitsdichte am Merkmalsvektor. Die Minimierung des Clusteringfehlers (3) ist die maximale Likelihood-Schätzung der Modellparameter (Bishop, 2006, Sect. 9.2). Der k-Means-Clusteringfehler ist ein Grenzwert dieser Wahl: Für gleiche Wahrscheinlichkeiten $p_c = 1/k$ und Kovarianzmatrizen $\mathbf{C}^{(c)} = \sigma^2\mathbf{I}$ ist der Verlust eines Datenpunkt im Clusteringfehler (3), wenn $\sigma \to 0$ und bis auf den Faktor $1/(2\sigma^2)$ und eine additive Konstante, der Verlust (2) (Bishop, 2006, Sect. 9.3.2).

Jeder dieser Clusteringfehler ist eine nicht-konvexe Funktion der Clusterzentren oder Modellparameter: mit einem einzelnen Datenpunkt bei $x = 0$ und zwei skalaren Clusterzentren $w^{(1)}, w^{(2)} \in \mathbb{R}$ ist der Clusteringfehler (1) $\min\{(w^{(1)})^2, (w^{(2)})^2\}$, der bei $(1, 0)$ und bei $(0, 1)$ verschwindet, aber nicht in deren Mittelpunkt. Der Lloyd-Algorithmus für k-Means (Lloyd, 1982) und der EM-Algorithmus (EM)-Algorithmus für ein GMM sind Fixpunktiterationen, die den Clusteringfehler nie erhöhen und an einem Fixpunkt anhalten, der ein lokales Minimum anstatt eines globalen sein kann (Bishop, 2006, Sect. 9.1). Sie werden daher von mehreren Initialisierungen aus gestartet, wobei der Lauf mit dem kleinsten Clusteringfehler beibehalten wird (Hastie et al., 2009, Sect. 14.3.6). Für die Tage in Krems senkt der Lloyd-Algorithmus, gestartet von den Mittelwerte der kälteren und der wärmeren Hälfte der Tage, den Clusteringfehler in drei Iterationen von 45.6 auf 44.8. Am Fixpunkt teilen sich 198 kalte Tage den Clusterzentrum $(3.9, 9.0)$ und 167 warme Tage den Clusterzentrum $(17.5, 24.9)$; jeder Tag im Januar und Februar ist ein kalter Tag, und jeder Tag im Juli und August ein warmer Tag.

Der Clusteringfehler kann die Anzahl k der Cluster nicht wählen. Ein zusätzlicher Clusterzentrum, der auf einem beliebigen Datenpunkt platziert wird, kann ihn nicht erhöhen, sodass der kleinste Clusteringfehler, der mit k Clusterzentren erreichbar ist, in k nicht zunimmt, und er ist null bei $k = m$, wo jeder Datenpunkt sein eigener Clusterzentrum ist. Dies ist der Trainingsfehler der Modellauswahl in anderer Form: Ein größerer Hypothesenraum hat nie einen größeren Trainingsfehler. Abb. 1(b) zeigt die Kurve für die Tage in Krems. Mit einem Clusterzentrum ist der Clusteringfehler die Summe der Varianzen der beiden Merkmale, 153.4; der zweite Clusterzentrum entfernt 108.5 davon, mehr als die nächsten sechs Clusterzentren zusammen (36.3). Ein solcher Knick in der Kurve ist der übliche informelle Leitfaden für k (Hastie et al., 2009, Sect. 14.3.11); andernfalls ist k ein Hyperparameter, der durch ein Kriterium, das ihn bestraft, oder durch die Leistung einer Vorhersage-Aufgabe, die die Cluster verwendet, gewählt wird.



(a) 21 of the 365 days and two Clusterzentren

(b) smallest clustering error found

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 den näheren von zwei Clusterzentren (Kreuze); der Clusteringfehler (1) mittelt die quadrierten Längen der gestrichelten Segmente über alle 365 Tage. (b) Der kleinste Clusteringfehler, gefunden mit $k = 1, \ldots, 8$ Clusterzentren: er fällt mit k und würde bei $k = 366$ Null erreichen. Daten generiert von `pythondemos/clusteringerror.py`

Siehe auch: Clustering, Cluster, Clusterzentrum, hartes Clustering, Soft Clustering, Zugehörigkeitsgrad, k-Means, Lloyd-Algorithmus, GMM, EM, maximale Likelihood, Verlust, Verlustfunktion, Trainingsfehler, ERM, euklidischer Abstand, Norm, GM, Median, Bregman-Divergenz, Ausreißer, Hyperparameter, Modellauswahl.

Querverweise

Literatur

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.