EM-Algorithmus (EM)

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 Expectation-Maximization (EM) Algorithmus ist ein iteratives Optimierungsverfahren für eine bestimmte Klasse von maximum likelihood-Problemen. Insbesondere ist der EM-Algorithmus nützlich für Wahrscheinlichkeitsmodelle, die beobachtete und unbeobachtete (oder latente) Attribute eines Datenpunkt umfassen. Jede Iteration des EM-Algorithmus wechselt einen E-Schritt mit einem M-Schritt ab. Der E-Schritt berechnet die A-posteriori-Verteilung der unbeobachteten Attribute unter den aktuellen Modellparameter. Der M-Schritt minimiert die negative Log-Likelihood der vollständigen Datenpunkte, gemittelt unter dieser A-posteriori-Verteilung, die für viele Wahrscheinlichkeitsmodelle eine geschlossene Form hat. Der EM- Algorithmus ist ein Beispiel für eine majorize–minimize (MM)-Methode. Eine kanonische soft clustering-Methode verwendet den EM- Algorithmus, um die Modellparameter eines Gaußsches Mischmodell (GMM) zu lernen.

Definition

Die Luft-Temperaturen, die um 06:00 (UTC) über ein Jahr an der Wetterstation Krems an der Donau (Österreich) aufgezeichnet wurden, streuen um zwei Regime: Morgenstunden der kalten und der warmen Jahreszeit. Ein Wahrscheinlichkeitsmodell für solche Datenpunkte hat neben der beobachteten Temperatur ein unbeobachtetes Attribut: das Regime, zu dem ein Morgen gehört. Abb. 1 zeigt das histogram der Temperaturen zusammen mit einem Gaußsches Mischmodell (GMM), das aus einer separaten Normalverteilung für jedes Regime besteht. Der EM-Algorithmus (Dempster et al., 1977) ist ein iteratives Optimierungsverfahren zur annähernden Lösung von maximum likelihood-Optimierungsprobleme für Wahrscheinlichkeitsmodelle mit einem unbeobachteten Attribut (Bishop, 2006, Abschn. 9.4; Hastie et al., 2009, Abschn. 8.5; Murphy, 2012, Abschn. 11.4.7).

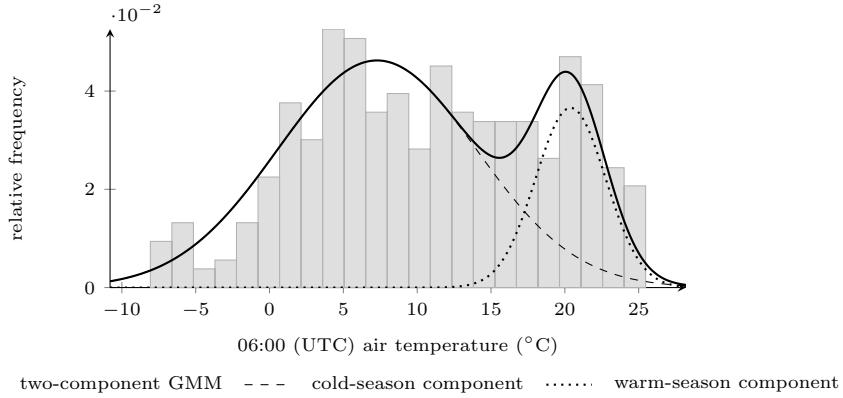


Abbildung 1: Histogram der 365 Lufttemperaturen, die um 06:00 (UTC) im Jahr 2024 an der Wetterstation Krems an der Donau aufgezeichnet wurden, mit dem zweikomponentigen GMM, das vom EM-Algorithmus gelernt wurde (durchgezogene Kurve), und seinen zwei gewichteten Komponenten, der Komponente der kalten Jahreszeit (gestrichelte Kurve) und der Komponente der warmen Jahreszeit (gepunktete Kurve). Jeder Morgen gehört zu einem der beiden Regime, aber welches, wird nicht beobachtet. Data generated by `pythondemos/em.py`

Man betrachte einen Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$, dessen Datenpunkte Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) sind. Die zugrunde liegende Wahrscheinlichkeitsverteilung ist ein Wahrscheinlichkeitsmodell $p^{(\mathbf{x}',y')}(\mathbf{x}, y; \mathbf{w})$ für den beobachteten Merkmalsvektor $\mathbf{x}$ und ein unbeobachtetes Attribut y mit Werten in einer endlichen Menge $\mathcal{Y}$. Dieses Wahrscheinlichkeitsmodell ist eine Familie von Wahrscheinlichkeitsverteilungen, parametrisiert durch Modellparameter $\mathbf{w}$ in einem parameter space $\mathcal{W}$. Die maximum likelihood ist ein Beispiel für empirische Risikominimierung (ERM): Sie minimiert das empirical risk

$$\widehat{L}\big(\mathbf{w}|\mathcal{D}\big) := -\frac{1}{m}\sum_{r=1}^{m} \ln \sum_{y\in\mathcal{Y}} p^{(\mathbf{x}',y')}\left(\mathbf{x}^{(r)}, y; \mathbf{w}\right), \tag{1}$$

die durchschnittliche negative Log-Likelihood der beobachteten Merkmalsvektoren. Die Summe innerhalb des Logarithmus macht es schwierig, (1) direkt zu minimieren: Wären die Attribute $y^{(r)}$ beobachtet, wäre das empirical risk ein Durchschnitt von $-\ln p^{(\mathbf{x}',y')}\left(\mathbf{x}^{(r)}, y^{(r)}; \mathbf{w}\right)$, was für viele Wahrscheinlichkeitsmodelle, darunter das GMM, einen Minimierer in geschlossener Form hat (Bishop, 2006, Abschn. 9.3.1).

Der EM-Algorithmus löst das maximum likelihood-Problem (1), indem er zwei Schritte abwechselt, die zusammen eine Iteration bilden. Der E-Schritt berechnet für jeden Datenpunkt die A-posteriori-Verteilung des unbeobachteten

Attributs unter den aktuellen Modellparameter $\widehat{\mathbf{w}}$,

$$p^{(y'|\mathbf{x}')}\left(y; \widehat{\mathbf{w}}\right) = \frac{p^{(\mathbf{x}',y')}\left(\mathbf{x}^{(r)}, y; \widehat{\mathbf{w}}\right)}{\sum_{y'' \in \mathcal{Y}} p^{(\mathbf{x}',y')}\left(\mathbf{x}^{(r)}, y''; \widehat{\mathbf{w}}\right)} \quad \text{für } y \in \mathcal{Y} \text{ und } r = 1, \ldots, m. \tag{2}$$

Diese A-posteriori-Verteilungen legen das Surrogat fest, das der E-Schritt von Algorithmus 1 bildet, das empirical risk der vollständigen Datenpunkte, gemittelt unter ihnen,

$$Q(\mathbf{w} \mid \widehat{\mathbf{w}}) := -\frac{1}{m} \sum_{r=1}^{m} \sum_{y \in \mathcal{Y}} p^{(y'|\mathbf{x}')}\left(y; \widehat{\mathbf{w}}\right) \ln \frac{p^{(\mathbf{x}',y')}\left(\mathbf{x}^{(r)}, y; \mathbf{w}\right)}{p^{(y'|\mathbf{x}')}\left(y; \widehat{\mathbf{w}}\right)}. \tag{3}$$

Der M-Schritt minimiert dieses Surrogat über $\mathbf{w} \in \mathcal{W}$. Algorithmus 1 beschreibt die Methode vollständig.

Algorithm 1 EM-Algorithmus (Bishop, 2006, Sect. 9.4)

Input: Datensatz $\mathcal{D}$, Wahrscheinlichkeitsmodell with Parameter $\mathbf{w} \in \mathcal{W}$, initialization $\mathbf{w}^{(0)}$, Iteration count T

1: **for** $t = 0, 1, \ldots, T-1$ **do**

2: **E-step:** form $Q\big(\mathbf{w} \mid \mathbf{w}^{(t)}\big)$, the Erwartungswert of the complete-data empirical risk under the current Parameter

3: **M-step:** $\mathbf{w}^{(t+1)} := \arg\min_{\mathbf{w} \in \mathcal{W}} Q\big(\mathbf{w} \mid \mathbf{w}^{(t)}\big)$

4: **end for**

Output: $\mathbf{w}^{(T)}$

Der EM-Algorithmus ist eine majorize–minimize (MM)-Methode: Jede Iteration minimiert ein Surrogat $Q(\mathbf{w} \mid \widehat{\mathbf{w}})$ des empirical risk (1). Das Surrogat (3) hat drei Eigenschaften (Bishop, 2006, Abschn. 9.4, dessen untere Schranke für die Log-Likelihood dieses Surrogat ist, geschrieben für die negative Log-Likelihood; Hunter and Lange, 2004, Abschn. 2). Erstens majorisiert es das empirical risk, $Q(\mathbf{w} \mid \widehat{\mathbf{w}}) \geq \widehat{L}\big(\mathbf{w}|\mathcal{D}\big)$ für alle $\mathbf{w} \in \mathcal{W}$, durch Jensens Ungleichung, angewendet auf den konkaven Logarithmus. Zweitens ist es am aktuellen Iterationspunkt scharf, $Q(\widehat{\mathbf{w}} \mid \widehat{\mathbf{w}}) = \widehat{L}\big(\widehat{\mathbf{w}}|\mathcal{D}\big)$. Drittens wird es durch den M-Schritt exakt minimiert. Der EM-Algorithmus ist keine coordinate descent auf dem empirical risk (1). Die coordinate descent minimiert das empirical risk selbst über jeweils einen Block der Modellparameter, während der E-Schritt keine Modellparameter variiert, sondern das Surrogat bildet, und der M-Schritt das Surrogat minimiert. Die beiden Schritte sind nur eine coordinate descent auf der erweiterten Zielfunktion, die die A-posteriori-Verteilungen des unbeobachteten Attributs als einen zweiten Variablenblock neben den Modellparameter behandelt (Bishop, 2006, Abschn. 9.4).

Beliebige Modellparameter $\mathbf{w}^{(t+1)}$, die das bei $\mathbf{w}^{(t)}$ gebildete Surrogat minimieren, erfüllen daher $\widehat{L}\big(\mathbf{w}^{(t+1)}|\mathcal{D}\big) \leq Q(\mathbf{w}^{(t+1)} \mid \mathbf{w}^{(t)}) \leq Q(\mathbf{w}^{(t)} \mid \mathbf{w}^{(t)}) = \widehat{L}\big(\mathbf{w}^{(t)}|\mathcal{D}\big)$: Keine Iteration erhöht das empirical risk (Dempster et al., 1977, Satz 1).

Die monotone Abnahme des empirical risk (1) entlang der EM-Iterationen führt zu einem stopping criterion: Die Iteration stoppt, sobald die durch die letzte Iteration erreichte Abnahme des empirical risk unter einen gegebenen Schwellenwert fällt (Bishop, 2006, Abschn. 9.2.2). Das empirical risk (1) ist im Allgemeinen eine nicht-konvexe Funktion der Modellparameter, und der EM-Algorithmus kann zu einem lokalen minimum konvergieren (Bishop, 2006, Abschn. 9.2.2). Er wird daher typischerweise mehrmals mit verschiedenen Initialisierungen ausgeführt, wobei der Lauf mit dem kleinsten empirical risk beibehalten wird (Murphy, 2012, Abschn. 11.3.2).

Der EM-Algorithmus ist eine fixed-point iteration $\widehat{\mathbf{w}}^{(t+1)} = \mathcal{F}\big(\widehat{\mathbf{w}}^{(t)}\big)$. Der Operator $\mathcal{F}$ bildet die aktuellen Modellparameter $\widehat{\mathbf{w}}$ auf das Ergebnis eines E-Schritts, gefolgt von einem M-Schritt, ab. Der E-Schritt (2) bei $\widehat{\mathbf{w}}$ bildet das Surrogat $Q(\cdot \mid \widehat{\mathbf{w}})$, und der M-Schritt (3) minimiert es,

$$\mathcal{F}(\widehat{\mathbf{w}}) := \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{W}} Q(\mathbf{w} \mid \widehat{\mathbf{w}}).$$

Die Fixpunkte von $\mathcal{F}$ sind Modellparameter, die ihr eigenes Surrogat minimieren: Dort berührt das Surrogat das empirical risk von oben, sodass kein weiterer Abstiegschritt möglich ist. Wenn das empirical risk und das Surrogat differentiable sind und der Fixpunkt $\widehat{\mathbf{w}}$ im Inneren von $\mathcal{W}$ liegt, verschwindet der Gradient des empirical risk bei $\widehat{\mathbf{w}}$. Tatsächlich ist die Differenz $Q(\mathbf{w} \mid \widehat{\mathbf{w}}) - \widehat{L}\big(\mathbf{w}|\mathcal{D}\big)$ für alle $\mathbf{w}$ nicht-negativ und null bei $\mathbf{w} = \widehat{\mathbf{w}}$, sodass ihr Gradient dort verschwindet, und der Gradient von $Q(\cdot \mid \widehat{\mathbf{w}})$ verschwindet bei seinem inneren minimum $\widehat{\mathbf{w}}$. Unter den Bedingungen von Wu (1983), Satz 2 ist jeder Häufungspunkt der EM-Iterationen ein solcher stationärer Punkt des empirical risk.
Der EM-Algorithmus wird verwendet, um die Modellparameter eines GMM zu lernen, wobei das unbeobachtete Attribut eines Datenpunkt sein Cluster-Index $I^{(r)} \in \{1, \ldots, k\}$ ist und die Modellparameter die Cluster-Wahrscheinlichkeiten p_c, Mittelwerte $\boldsymbol{\mu}^{(c)}$ und Kovarianzmatrizen $\mathbf{C}^{(c)}$ sind (Bishop, 2006, Sect. 9.2.2). Der E-Schritt (2) wird zur Zugehörigkeitsgrad

$$u_c^{(r)} := \mathbb{P}\left(I^{(r)} = c \mid \mathbf{x}^{(r)}\right) = \frac{p_c \, \mathcal{N}\left(\boldsymbol{\mu}^{(c)}, \mathbf{C}^{(c)}\right)\left(\mathbf{x}^{(r)}\right)}{\sum_{c'=1}^{k} p_{c'} \, \mathcal{N}\left(\boldsymbol{\mu}^{(c')}, \mathbf{C}^{(c')}\right)\left(\mathbf{x}^{(r)}\right)}, \tag{4}$$

und die Minimierung des Surrogats (3) liefert den M-Schritt in geschlossener Form. Die Cluster-Wahrscheinlichkeit p_c ist der Durchschnitt von $u_c^{(r)}$ über die Datenpunkte, während $\boldsymbol{\mu}^{(c)}$ und $\mathbf{C}^{(c)}$ der Mittelwert und die Kovarianzmatrix der Merkmalsvektoren sind, gewichtet mit $u_c^{(r)}$. Die gelernten Zugehörigkeitsgrade (4) bilden eine soft clustering. Wenn jede Cluster-Wahrscheinlichkeit auf $1/k$ und jede Kovarianzmatrix auf $\sigma^2\mathbf{I}$ festgelegt wird, ein Vielfaches der Einheits-Matrix $\mathbf{I}$ mit einem gemeinsamen σ^2, das gegen 0 schrumpft, werden die Zugehörigkeitsgrade zu 0 oder 1, und der EM-Algorithmus reduziert sich auf den Lloyd's algorithm für k-Means (Bishop, 2006, Sect. 9.3.2).

Betrachten wir als Nächstes, zur Veranschaulichung des EM-Algorithmus, das GMM von Abb. 1 für den Temperatur-Datensatz. Das Merkmal eines Da-

tenpunkt ist der 06:00 (UTC) Messwert $x \in \mathbb{R}$, und sein unbeobachtetes Attribut ist das Regime $y \in \{\text{cold}, \text{warm}\}$. Die Modellparameter sind die Wahrscheinlichkeiten p_{cold}, $p_{\text{warm}} = 1 - p_{\text{cold}}$ der beiden Regime, die Mittelwerte μ_{cold}, μ_{warm} und die Varianzen σ^2_{cold}, σ^2_{warm} der beiden Normalverteilungen, so $\mathbf{w} = (p_{\text{cold}}, p_{\text{warm}}, \mu_{\text{cold}}, \mu_{\text{warm}}, \sigma^2_{\text{cold}}, \sigma^2_{\text{warm}})^\top \in \mathbb{R}^6$. Der E-Schritt (2) berechnet für jeden Morgen die Wahrscheinlichkeit des kalten Regimes, und der M-Schritt (3) berechnet $\mathbf{w}$ als Durchschnitte neu, gewichtet mit diesen Wahrscheinlichkeiten. An den 365 Morgen des Jahres 2024 lernt der EM-Algorithmus $p_{\text{cold}} = 0.78$, $\mu_{\text{cold}} = 7.3$, $\mu_{\text{warm}} = 20.4$, $\sigma^2_{\text{cold}} = 45.6$ und $\sigma^2_{\text{warm}} = 5.6$.

Abb. 2 zeigt die MM-Ansicht für dieses GMM, während der Mittelwert der kalten Jahreszeit μ_{cold} variiert, während p_{cold}, μ_{warm} und σ^2_{warm} auf ihren gelernten Werten und σ^2_{cold} auf 16 fixiert sind. Das empirical risk ist dann eine Funktion von μ_{cold} allein, und so auch jedes Surrogat. Jedes Surrogat liegt über dem empirical risk, berührt es an seinem Iterat, und sein Minimierer ist das nächste Iterat. Abb. 3 zeigt, wie das empirical risk monoton abnimmt, während die EM-Iterationen fortschreiten.

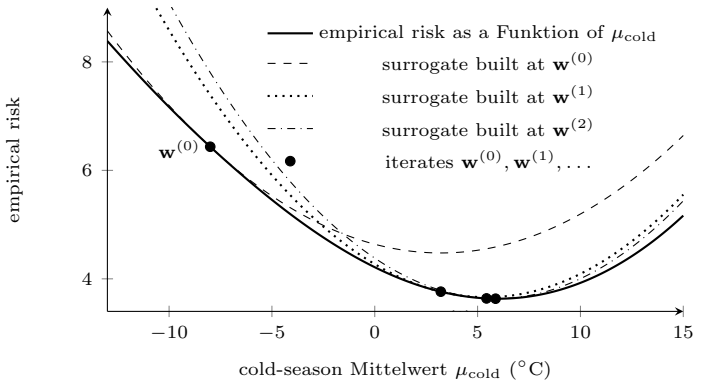


Abbildung 2: Die MM-Ansicht des EM-Algorithmus für das zweikomponentige GMM von Abb. 1 mit einem einzigen freien Modellparameter, dem Mittelwert der kalten Jahreszeit μ_{cold}. Die anderen Modellparameter sind fixiert: p_{cold}, μ_{warm} und σ^2_{warm} auf ihren gelernten Werten, und $\sigma^2_{\text{cold}} = 16$, kleiner als sein gelernter Wert, damit sich die Surrogate sichtbar vom empirical risk unterscheiden. Durchgezogen: das empirical risk als Funktion von μ_{cold}. Gestrichelt, gepunktet und strichpunktiert: die Surrogate, die vom E-Schritt bei den ersten drei Iteraten $\mathbf{w}^{(0)}, \mathbf{w}^{(1)}, \mathbf{w}^{(2)}$ des EM-Algorithmus auf μ_{cold} gebildet wurden, beginnend bei $\mu_{\text{cold}} = -8$. Jedes Surrogat liegt über dem empirical risk, berührt es an seinem Iterat und wird beim nächsten Iterat minimiert. Data generated by `pythondemos/em.py`

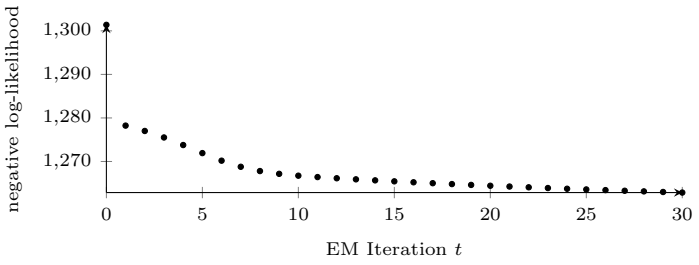


Abbildung 3: Der EM-Algorithmus lernt alle sechs Modellparameter des zweikomponentigen GMM von Abb. 1 anhand der 365 Morgenmessungen von 2024 in Krems: die negative Log-Likelihood nimmt monoton ab, während die Iterationen fortschreiten, und stabilisiert sich an einem Fixpunkt des EM-Operator $\mathcal{F}$. Data generated by `pythondemos/em.py`

Siehe auch: maximum likelihood, Wahrscheinlichkeitsmodell, ERM, MM, fixed-point iteration, Fixpunkt, A-posteriori-Verteilung, GMM, soft clustering, Zugehörigkeitsgrad, k-Means, Lloyd's algorithm, coordinate descent, Optimierungsproblem.

Querverweise

Literatur

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

2. Hunter and Lange (2004). A Tutorial on MM Algorithms. Amer. Statistician, 58(1), pp. 30–37.

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

4. Wu (1983). On the convergence properties of the EM algorithm. Ann. Statist., 11(1), pp. 95–103.

5. 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.

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.