maximale Likelihood

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

Die maximale Likelihood wählt diejenigen Modellparameter eines Wahrscheinlichkeitsmodells, unter denen ein beobachteter Datensatz die größte Wahrscheinlichkeitsdichte (oder Wahrscheinlichkeitsmasse) besitzt, also die Modellparameter, welche die Likelihood-Funktion maximieren. Der Datensatz wird als Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) mit einer gemeinsamen Wahrscheinlichkeitsdichte aus einer durch die Modellparameter indizierten Familie modelliert. Durch den negativen Logarithmus geht das Produkt der Wahrscheinlichkeitsdichten in ein Mittel von Verluste über, sodass die maximale Likelihood der empirische Risikominimierung (ERM) mit der negativen Log-Likelihood als Verlustfunktion entspricht; die lineare Regression mit gaußschem Rauschen, die logistische Regression und das Gaußsches Mischmodell (GMM) sind maximale Likelihood unter spezifischen Wahrscheinlichkeitsmodelleen. Für eine Normalverteilung sind die Schätzwerte das Stichprobenmittel und der Mittelwert der quadratischen Abweichung; wo keine geschlossene Form existiert, berechnet der Gradientenabstieg oder der EM-Algorithmus (EM)-Algorithmus den Schätzwert als Fixpunktiteration. Unter Regularitätsbedingungen konvergiert der Schätzwert mit wachsendem Datensatz gegen die wahren Modellparameter und ist asymptotisch normalverteilt mit der Varianz der Cramér-Rao-Schranke. Bei einem großen Hypothesenraum überanpasst der Schätzwert, und das Hinzufügen eines Strafterms zur negativen Log-Likelihood, was einer Maximum-a-posteriori-Schätzung entspricht, oder das Verkleinern des Hypothesenraums regularisiert ihn.

Definition

Betrachtet seien die 365 Lufttemperaturen, die im Jahr 2024 jeweils um 15:00 Uhr (UTC) an der Wetterstation Krems an der Donau (Österreich) aufgezeichnet wurden, ein Wert pro Tag, den die stündliche Aufzeichnung abdeckt. Abb. 1(a) zeigt ihr Histogramm. Ein Wahrscheinlichkeitsmodell für diese Zahlen behandelt jede von ihnen als Realisierung einer Zufallsvariable (ZV), deren Wahrscheinlichkeitsdichte zu einer durch wenige Modellparameter indizierten Familie gehört, wie etwa den durch Mittelwert und Varianz indizierten Normalverteilungen. Die maximale Likelihood wählt diejenigen Modellparameter, unter denen das aufgezeichnete Jahr am wahrscheinlichsten ist: Unter allen Normal-

verteilungen greift sie diejenige heraus, deren Wahrscheinlichkeitsdichte, ausgewertet an den 365 beobachteten Temperaturen und ausmultipliziert, am größten ist. Für die Tage in Krems ist dies die Normalverteilung mit dem Mittelwert 16.3°C und der Standardabweichung 9.3°C, eingezeichnet in Abb. 1(a).

Formal sei ein Datensatz $\mathcal{D} = \{\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}\}$ betrachtet, dessen Datenpunkte Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen mit einer gemeinsamen Wahrscheinlichkeitsdichte (oder Wahrscheinlichkeitsfunktion (pmf)) $p^{(\mathbf{z};\mathbf{w})}(\cdot)$ sind, die von den Modellparametern $\mathbf{w} \in \mathcal{W}$ abhängt. Die Likelihood-Funktion von $\mathbf{w}$ ist die gemeinsame Wahrscheinlichkeitsdichte des Datensatzes, aufgefasst als Funktion von $\mathbf{w}$ bei festgehaltenem Datensatz,

$$L^{(\mathcal{D})}(\mathbf{w}) := \prod_{r=1}^{m} p^{(\mathbf{z};\mathbf{w})}\left(\mathbf{z}^{(r)}\right). \tag{1}$$

Der Schätzwert der maximalen Likelihood ist ein Maximierer der Likelihood-Funktion (1) (Casella and Berger, 2002; Fisher, 1922; Lehmann and Casella, 1998, Def. 3.1, S. 444). Da der Logarithmus monoton wachsend ist, ist er zugleich ein Minimierer der negativen Log-Likelihood pro Datenpunkt,

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{m} \sum_{r=1}^{m} \big(-\ln p^{(\mathbf{z};\mathbf{w})}\left(\mathbf{z}^{(r)}\right)\big), \tag{2}$$

worin das Produkt zu einer Summe geworden ist. Die Wahrscheinlichkeitsdichte wird bezüglich eines Basis-Maßes gebildet, das nicht von $\mathbf{w}$ abhängt; ein Wechsel dieses Maßes multipliziert die Likelihood-Funktion mit einem von $\mathbf{w}$ unabhängigen Faktor, sodass der Schätzwert unverändert bleibt (Hastie et al., 2009, Abschn. 8.2.2).

Die negative Log-Likelihood pro Datenpunkt im Schätzwert (2) ist ein mittlerer Verlust. Jeder Wert $\mathbf{w}$ der Modellparameter definiert eine Hypothese $h^{(\mathbf{w})}$, die Wahrscheinlichkeitsdichte $p^{(\mathbf{z};\mathbf{w})}(\cdot)$, und der Verlust, den sie auf einem Datenpunkt $\mathbf{z}$ verursacht, ist $L\left(\mathbf{z}, h^{(\mathbf{w})}\right) := -\ln p^{(\mathbf{z};\mathbf{w})}(\mathbf{z})$. Ein Datenpunkt, an dem die Wahrscheinlichkeitsdichte klein ist, kostet viel, einer, an dem sie groß ist, kostet wenig. Die maximale Likelihood ist daher empirische Risikominimierung (ERM) über dem Hypothesenraum $\mathcal{H} = \{h^{(\mathbf{w})} : \mathbf{w} \in \mathcal{W}\}$ mit der negativen Log-Likelihood als Verlustfunktion, und das Minimum im Schätzwert (2) ist der Trainingsfehler der gelernten Hypothese $\widehat{h} = h^{(\widehat{\mathbf{w}})}$. Mehrere Methoden des maschinelles Lernen (ML)s sind maximale Likelihood unter einem spezifischen Wahrscheinlichkeitsmodell. Die lineare Regression mit gaußschem Rauschen fester Varianz ist maximale Likelihood mit dem quadratischer Verlust (Bishop, 2006, Abschn. 3.1.1). Die logistische Regression ist maximale Likelihood mit dem logistische Verlustfunktion, der negativen Log-Likelihood des Labels unter einem Wahrscheinlichkeitsmodell, in dem die Sigmoidfunktion einer linearen Funktion die Wahrscheinlichkeit des Labels ist (Bishop, 2006, Abschn. 4.3.2). Das Lernen eines Gaußsches Mischmodell (GMM)s ist maximale Likelihood mit dem Clusteringfehler eines Soft Clusterings, dem negativen Logarithmus der Mischungs-Wahrscheinlichkeitsdichte (Bishop, 2006, Abschn. 9.2.1).

Betrachtet sei die Normalverteilung $\mathcal{N}\left(\mu, \sigma^2\right)$ für skalare Datenpunkte $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}$, sodass die Modellparameter durch $\mathbf{w} = (\mu, \sigma^2)$ gegeben sind. Die negative Log-Likelihood pro Datenpunkt im Schätzwert (2) lautet dann

$$\frac{1}{2} \ln\left(2\pi\sigma^2\right) + \frac{1}{2\sigma^2 m} \sum_{r=1}^{m} \left(x^{(r)} - \mu\right)^2. \tag{3}$$

Für ein festes σ^2 ist die gaußsche negative Log-Likelihood (3) eine streng monoton wachsende affine Funktion des mittleren quadratischen Fehlers, sodass der Mittelwert mit der größten Likelihood-Funktion der Mittelwert mit dem kleinsten mittleren quadratischen Fehler ist. Die Nullgradientenbedingung für die gaußsche negative Log-Likelihood (3) liefert die geschlossene Form

$$\widehat{\mu} = \frac{1}{m} \sum_{r=1}^{m} x^{(r)}, \qquad \widehat{\sigma}^2 = \frac{1}{m} \sum_{r=1}^{m} \left(x^{(r)} - \widehat{\mu}\right)^2, \tag{4}$$

das Stichprobenmittel und den Mittelwert der quadratischen Abweichung davon (Bishop, 2006, Abschn. 1.2.4, Gl. (1.55)–(1.56)). Für die Tage in Krems gilt $\widehat{\mu} = 16.3°\mathrm{C}$ und $\widehat{\sigma}^2 = 87.2$ Quadratgrad, und kein Paar von Modellparametern auf einem feinen Gitter um diese Werte hat eine größere Likelihood-Funktion. Abb. 1(b) zeigt die mittlere Log-Likelihood als Funktion von μ bei auf $\widehat{\sigma}^2$ festgehaltenem σ^2: eine nach unten geöffnete Parabel mit ihrem Scheitelpunkt beim Stichprobenmittel. Der Schätzwert der Varianz in der geschlossenen Form (4) ist verzerrt. Sein Erwartungswert beträgt $(m-1)\sigma^2/m$, sodass er die Varianz um den Faktor $(m-1)/m$ unterschätzt, der für die 365 Tage bei 0.997 liegt (Bishop, 2006, Abschn. 1.2.4, Gl. (1.58)). Die geschlossene Form wird in einem Schritt berechnet, sodass keine Iteration vorliegt, die als Fixpunktiteration gedeutet werden könnte.

Für ein GMM, für die logistische Regression und für die meisten Wahrscheinlichkeitsmodellee mit mehr als einigen wenigen Modellparametern besitzt die Nullgradientenbedingung keine geschlossene Lösung, und der Schätzwert (2) wird mit einem iterativen Verfahren berechnet. Der Gradientenabstieg auf der mittleren negativen Log-Likelihood ist eine Fixpunktiteration $\mathbf{w}^{(t+1)} = \mathcal{F}\left(\mathbf{w}^{(t)}\right)$ für $t = 0, 1, \ldots$, deren Fixpunkte die Punkte sind, an denen der Gradient der negativen Log-Likelihood verschwindet. Für Wahrscheinlichkeitsmodellee mit einem unbeobachteten Attribut, wie etwa dem Cluster-Index eines GMMs, ist der EM-Algorithmus (EM)-Algorithmus (Dempster et al., 1977) die Standard-Fixpunktiteration: Eine Iteration verringert die Likelihood-Funktion nie, und ihre Fixpunkte sind wiederum Punkte mit verschwindendem Gradienten (Bishop, 2006, Abschn. 9.2.2). Die negative Log-Likelihood eines GMMs ist nicht konvex, sodass jedes der beiden Verfahren in einem lokalen Minimum enden kann (Bishop, 2006, Abschn. 9.2.2).

Wie nahe liegt $\widehat{\mathbf{w}}$ an den Modellparametern $\mathbf{w}^\star$ der Wahrscheinlichkeitsverteilung, die den Datensatz erzeugt hat? Dies ist die Frage nach der Generalisierung bei der maximalen Likelihood, gestellt unter der Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme). Der Erwartungswert des Ver-

lusts $L\left(\mathbf{z}, h^{(\mathbf{w})}\right)$ unter $p^{(\mathbf{z};\mathbf{w}^\star)}(\cdot)$ ist das Risiko von $h^{(\mathbf{w})}$. Es übersteigt das Risiko von $h^{(\mathbf{w}^\star)}$ um die Kullback-Leibler-Divergenz (KL-Divergenz) von $p^{(\mathbf{z};\mathbf{w}^\star)}(\cdot)$ zu $p^{(\mathbf{z};\mathbf{w})}(\cdot)$, die nichtnegativ ist (Bishop, 2006, Sect. 1.6.1). Die wahren Modellparameter minimieren daher das Risiko, und die maximale Likelihood minimiert sein empirisches Gegenstück. Betrachtet sei ein einzelner Modellparameter $w \in \mathcal{W} \subseteq \mathbb{R}$ mit einem wahren Wert $w^\star$ im Inneren von $\mathcal{W}$. Angenommen sei, dass die Wahrscheinlichkeitsdichte glatt in w ist, dass verschiedene Werte von w verschiedene Wahrscheinlichkeitsverteilungen ergeben und dass der Träger der Wahrscheinlichkeitsdichte nicht von w abhängt. Dann besitzt die Nullgradientenbedingung der Log-Likelihood, mit einer gegen eins strebenden Wahrscheinlichkeit für $m \to \infty$, eine Nullstelle $\widehat{w}$, die in Wahrscheinlichkeit gegen $w^\star$ konvergiert (Lehmann and Casella, 1998, Thm. 3.7, p. 447). Die maximale Likelihood ist konsistent. Jede solche konsistente Folge von Nullstellen ist asymptotisch normalverteilt: Mit der Fisher-Information $I(w) := \mathbb{E}\left\{\left(\frac{\partial}{\partial w} \ln p^{(\mathbf{z};w)}(\mathbf{z})\right)^2\right\}$ eines Datenpunktes gilt

$$\sqrt{m}\left(\widehat{w} - w^\star\right) \to \mathcal{N}\left(0, 1/I(w^\star)\right) \quad \text{in Verteilung für } m \to \infty, \tag{5}$$

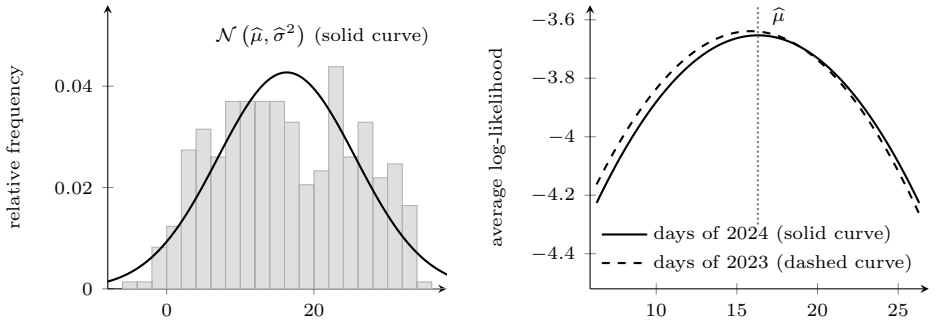
(Hastie et al., 2009, Sect. 8.2.2, Eq. (8.16); Lehmann and Casella, 1998, Thm. 3.10, p. 449). Die Informationsungleichung, auch bekannt als Cramér-Rao-Schranke, besagt, dass kein unverzerrter Schätzer von w, der aus m Datenpunkteen berechnet wird und milde Regularitätsbedingungen erfüllt, eine Varianz unterhalb von $1/(mI(w))$ aufweist (Cramér, 1946; Lehmann and Casella, 1998, Thm. 5.10, p. 120). Die asymptotische Normalität (5) besagt, dass sich die Varianz von $\widehat{w}$ dieser Schranke annähert, weshalb die maximale Likelihood als asymptotisch effizient bezeichnet wird. Für den Mittelwert einer Normalverteilung mit bekannter Varianz σ^2 beträgt die Fisher-Information $I(\mu) = 1/\sigma^2$, und das Stichprobenmittel erreicht die Schranke exakt: Seine Varianz ist σ^2/m. Für die Tage in Krems beträgt die Standardabweichung von $\widehat{\mu}$ daher $\widehat{\sigma}/\sqrt{365} = 0.49°\mathrm{C}$. Die zweite Ableitung der Parabel in Abb. 1(b) ist $-1/\widehat{\sigma}^2$, die Fisher-Information mit negativem Vorzeichen. Die 364 Tage des Jahres 2023 ergeben im selben Teilbild eine zweite Parabel, deren Scheitelpunkt beim Mittelwert von 2023 bei 15.9°C liegt, vier Zehntelgrad unter dem Schätzwert für 2024.

Die maximale Likelihood wählt diejenigen Modellparameter, die für die Trainingsdatensatz am besten sind, und bei einem großen Hypothesenraum führt dies zu Überanpassung (Murphy, 2012, Sect. 7.5.1). Eine Vergrößerung des Hypothesenraums für die Tage in Krems zeigt diesen Effekt. Mit zwölf Mittelwerteen, einem pro Kalendermonat, und einer gemeinsamen Varianz steigt die durchschnittliche Log-Likelihood auf den 365 Tagen des Jahres 2024 von -3.65 auf -2.87. Sie steigt auch auf den 364 Tagen des Jahres 2023, die nicht in die Schätzung eingeflossen sind, von -3.64 auf -3.12: Die Jahreszeiten sind real, sodass der größere Hypothesenraum auch auf neuen Daten hilft. Mit einem Mittelwert pro Kalendertag legt die maximale Likelihood jeden Mittelwert genau auf seine Beobachtung, und der Schätzwert für die Varianz ist null. Die Likelihood-Funktion von 2024 übersteigt dann für $\sigma \to 0$ jede Schranke,

während die durchschnittliche Log-Likelihood von 2023 unter denselben Modellparametern unter jede Schranke fällt: -23 bei $\sigma = 1°\mathrm{C}$ und -2237 bei $\sigma = 0.1°\mathrm{C}$. Dieselbe Singularität tritt in einem GMM mit einer Komponente auf, deren Mittelwert mit einem Datenpunkt zusammenfällt und deren Varianz gegen null schrumpft (Bishop, 2006, Sect. 9.2.1). Beide Gegenmaßnahmen sind Formen der Regularisierung. Die eine verkleinert den Hypothesenraum: weniger Mittelwertee oder eine untere Schranke für die Varianz. Die andere fügt der negativen Log-Likelihood einen Strafterm hinzu. Dies ist die Maximum-a-posteriori-Schätzung (MAP-Schätzung),

$$\widehat{\mathbf{w}} \in \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}} \frac{1}{m} \sum_{r=1}^{m} \big(- \ln p^{(\mathbf{z};\mathbf{w})} \left(\mathbf{z}^{(r)} \right) \big) - \frac{1}{m} \ln p^{(\mathbf{w})} \left(\mathbf{w} \right), \tag{6}$$

bei der der Strafterm der negative Logarithmus einer Prior-Verteilung mit der Wahrscheinlichkeitsdichte $p^{(\mathbf{w})}(\cdot)$ über den Modellparametern ist. Ein Minimierer der bestraften negativen Log-Likelihood (6) maximiert die A-posteriori-Verteilung von $\mathbf{w}$ gegeben den Datensatz (Bishop, 2006, Sect. 1.2.5, Eq. (1.67)). Eine Gaußsche Prior-Verteilung auf dem Gewichts-Vektor der lineare Regression ergibt die Ridge-Regression (Murphy, 2012, Sect. 7.5.1). Eine Gaußsche Prior-Verteilung $\mathcal{N}\left(\mu_0, \tau^2\right)$ auf dem Mittelwert einer Normalverteilung mit bekanntem σ^2 verschiebt den Schätzwert vom Stichprobenmittel $\widehat{\mu}$ in Richtung μ_0 auf $(m\tau^2\widehat{\mu} + \sigma^2\mu_0)/(m\tau^2 + \sigma^2)$ (Bishop, 2006, Sect. 2.3.6, Eq. (2.141)). Der Strafterm im MAP-Schätzwert (6) wird durch m geteilt, sodass er mit wachsendem Datensatz verschwindet und sich der MAP-Schätzwert dem Schätzwert der maximalen Likelihood annähert.



(a) Histogramm and maximum likelihood Wahrscheinlichkeitsdichte (b) average log-likelihood against the Mittelwert

Abbildung 1: (a) Histogramm der 365 Lufttemperaturen, gemessen um 15:00 Uhr (UTC) im Jahr 2024 in Krems an der Donau (Balken, Klassenbreite 2°C), sowie die Wahrscheinlichkeitsdichte der Normalverteilung mit der größten Likelihood-Funktion (durchgezogene Kurve), mit Mittelwert $\widehat{\mu} = 16.3$°C und Standardabweichung $\widehat{\sigma} = 9.3$°C. Die Jahreszeiten verteilen die Messwerte über 40 Grad mit einem flachen Plateau und einem Einbruch nahe 19°C, dem keine einzelne Normalverteilung folgt. (b) Die durchschnittliche Log-Likelihood der 365 Tage von 2024 (durchgezogene Kurve) und der 364 Tage von 2023 (gestrichelte Kurve) als Funktion des Mittelwerts μ, bei festgesetzter Varianz $\widehat{\sigma}^2$; die gepunktete Linie markiert den Schätzwert der maximalen Likelihood $\widehat{\mu}$ von 2024. Daten erzeugt mit `pythondemos/maxlikelihood.py`

Siehe auch: Likelihood-Funktion, Wahrscheinlichkeitsmodell, Wahrscheinlichkeitsverteilung, Wahrscheinlichkeitsdichte, Modellparameter, Schätzer, ERM, Verlustfunktion, Trainingsfehler, Risiko, u.i.v.-Annahme, Normalverteilung, Stichprobenmittel, Varianz, lineare Regression, quadratischer Verlust, logistische Regression, logistische Verlustfunktion, GMM, Clusteringfehler, EM, Gradientenabstieg, Fixpunktiteration, Nullgradientenbedingung, KL-Divergenz, Überanpassung, Regularisierung, Strafterm, Prior-Verteilung, A-posteriori-Verteilung, Ridge-Regression, Optimierungsproblem.

Querverweise

Literatur

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

2. Casella and Berger (2002). Statistical Inference. 2nd ed., Duxbury, Pacific Grove, CA.

3. Fisher (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London. Series A, 222, pp. 309–368.

4. Lehmann and Casella (1998). Theory of Point Estimation. 2nd ed., Springer-Verlag, New York, NY, USA.

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

6. Cramér (1946). A contribution to the theory of statistical estimation. Skand. Akt. Tidskr., 29, pp. 85–94.

7. 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.

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