Baseline

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

Eine Baseline ist ein Referenzniveau, mit dem die Leistung eines trainierten Modells verglichen wird. Die einfachste Baseline verwendet überhaupt keine Merkmale und sagt für jeden Datenpunkt denselben Wert voraus, z. B. den Stichprobenmittel numerischer Label-Werte oder den häufigsten Label-Wert bei der Klassifikation. Jede nützliche Methode muss dieses Niveau übertreffen. Baselines lassen sich aus menschlicher Leistung, aus einem bestehenden maschinelles Lernen (ML)-Verfahren oder aus einem Wahrscheinlichkeitsmodell gewinnen. Ist ein genaues Wahrscheinlichkeitsmodell der Datenerzeugung gegeben, so ist das kleinste erreichbare Risiko das Bayes-Risiko, das durch den Bayes-Schätzer des Labels bei gegebenen Merkmalen entsteht. In der Praxis ist der Bayes-Schätzer jedoch aus mindestens zwei Gründen nicht realisierbar. Erstens lässt sich die Wahrscheinlichkeitsverteilung des Datenerzeugungsprozesses möglicherweise nicht bestimmen. Zweitens kann die Berechnung des Bayes-Risikos rechnerisch zu aufwendig sein.

Definition

An drei Tagen des Jahres 2024 verzeichnete die Wetterstation Krems an der Donau (Österreich) um 06:00 Uhr (UTC) dieselbe Lufttemperatur von 0.7 °C, um 15:00 Uhr (UTC) jedoch einen Messwert von 1.0 °C an einem Tag, 5.8 °C am zweiten und 17.3 °C am dritten. Keine Hypothese, die den Nachmittagswert aus dem Morgenwert vorhersagt, kann an allen drei Tagen richtig liegen. Welchen Verlust eine Hypothese an diesen drei Tagen auch erleidet, er ist unvermeidbar; ein gemessener Verlust über null ist daher für sich genommen noch kein Zeichen für ein schlechtes ML-Verfahren.

Allgemeiner sei ein ML-Verfahren betrachtet, das eine gelernte Hypothese (oder ein trainiertes Modell) $\widehat{h} \in \mathcal{H}$ erzeugt. Es wird bewertet, indem der durchschnittliche Verlust auf einer Testdatensatz berechnet wird (Hastie et al., 2009, Sect. 7.2). Die Frage ist, ob die gelernte Hypothese bereits nahezu optimal ist. Ist dies der Fall, bringt das Sammeln weiterer Daten oder zusätzlicher Rechenaufwand zu ihrer Verbesserung nur wenig Gewinn. Um diese Frage zu beantworten, wird eine Baseline benötigt: ein Referenzniveau für den durchschnittlichen Verlust, entweder das kleinste, das überhaupt ein Verfahren erreichen kann, oder das Niveau, das jedes nützliche Verfahren unterbieten muss.

Manchmal wird eine Baseline aus dem Datensatz selbst berechnet, wie im Streudiagramm in Abb. 1 (Hastie et al., 2009, Sect. 2.4). Die Abbildung zeigt Wetteraufzeichnungen der Station Krems an der Donau (Österreich). Jeder der 365 Tage des Jahres 2024, für die beide Messwerte vorliegen, ist ein Datenpunkt, dessen Merkmal x der Messwert des Tages um 06:00 Uhr (UTC) und dessen Label y sein Messwert um 15:00 Uhr (UTC) ist. Für die 11 Datenpunkte, deren x zwischen 11.5 °C und 12.5 °C liegt, reichte das zugehörige Label von 12.4 °C bis 28.4 °C. Drei Datenpunkte des Jahres teilen sogar denselben Merkmalswert $x = 0.7$, mit den Labelsn 1.0, 5.8 und 17.3. Da eine Hypothese eine Abbildung ist, weist sie allen drei Tagen dieselbe Vorhersage zu. Auf diesen drei Datenpunkten erleidet sie daher einen durchschnittlichen quadratischen Fehler von mindestens 46.8, der Stichprobenvarianz dieser drei Labels. Dasselbe gilt überall dort, wo Datenpunkte denselben Merkmalswert teilen: Die Streuung ihrer Labels bestimmt einen durchschnittlichen Verlust, den keine Hypothese, wie auch immer sie gelernt wurde, vermeiden kann.

Die grauen Streifen in Abb. 1 fassen Datenpunkte mit sehr ähnlichen Merkmalswerten zusammen. Jeder Streifen enthält mindestens vier Datenpunkte und erstreckt sich über die Spannweite der Labels, die diese Datenpunkte tragen. Ein Streifen ist daher dort schmal, wo die Datenpunkte entlang der Merkmalsachse dicht liegen, und dort breit, wo sie dünn besetzt sind; und er ist dort hoch, wo diese Datenpunkte sehr unterschiedliche Labels tragen. Berechnet man die Stichproben-Varianz der Labels innerhalb jedes Streifens und mittelt die Ergebnisse über alle Streifen, erhält man 14.69 (in Grad Celsius zum Quadrat). Der resultierende Wert von 14.69 Grad Celsius zum Quadrat ist eine Schätzung des kleinsten durchschnittlichen quadratischen Fehlers, den irgendeine Hypothese auf diesen Datenpunkten erreichen kann (Hastie et al., 2009, Sect. 2.4). Die mittels linearer Regression gelernte Hypothese, die durchgezogene Linie in Abb. 1, erleidet einen durchschnittlichen quadratischen Fehler von 15.25, ihren Trainingsfehler. Der Trainingsfehler liegt innerhalb von vier Prozent der Baseline, ein Beleg dafür, dass kaum noch eine Verbesserung zu erzielen ist.

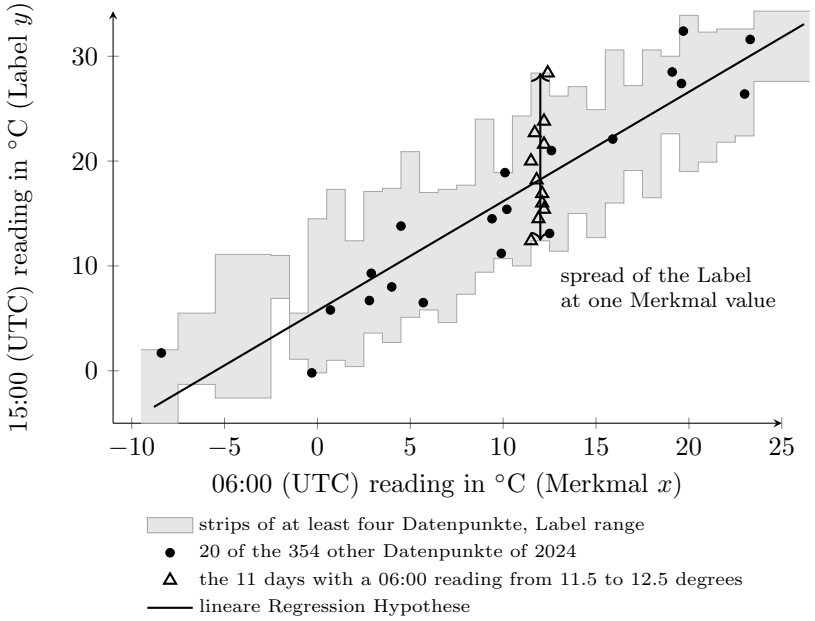

Abbildung 1: Beobachtungen, die im Jahr 2024 an der Wetterstation Krems an der Donau gesammelt wurden. Jeder Datenpunkt ist ein Tag, mit dem Messwert um 06:00 (UTC) als seinem Merkmal x und dem Messwert um 15:00 (UTC) als seinem Label y. Die grau schattierten Streifen liefern die aus den Daten abgeleitete Baseline von 14.69 Quadratgrad Celsius. Die grauen Streifen unterteilen die Merkmal-Achse in Abschnitte, die jeweils mindestens vier Datenpunkte enthalten und jeweils über die Spannweite der Labels ihrer Datenpunkte gezeichnet sind; ihre Breiten unterscheiden sich, da die Datenpunkte nicht gleichmäßig über die Merkmal-Achse verteilt sind. Die offenen Dreiecke markieren die 11 Datenpunkte eines solchen Streifens, nämlich desjenigen um $x = 12$, dessen Merkmal-Werte in $[11.5, 12.5)$ liegen, und der Doppelpfeil überspannt die Labels, die diese 11 Datenpunkte tragen. Die ausgefüllten Kreise sind 20 der 354 Datenpunkte außerhalb dieses Streifens, nämlich jeder 18. davon. Die durchgezogene Linie ist die durch lineare Regression gelernte Hypothese. Gemittelt über alle Streifen schätzt eine solche Streuung den kleinsten erreichbaren durchschnittlichen Verlust für jede Methode des maschinellen Lernens, die den Messwert um 15:00 ausschließlich aus demjenigen um 06:00 vorhersagt. Daten generiert mit `pythondemos/baseline.py`

Baselines können untere oder obere Schranken für den erreichbaren durchschnittlichen Verlust liefern. Die streifenweise Schätzung aus Abb. 1 approximiert eine untere Schranke, das unten definierte Bayes-Risiko: Keine Hypothese hat ein kleineres Risiko.

Eine obere Schranke, d. h. eine Baseline, die jedes nützliche Verfahren schla-

gen sollte, ergibt sich aus einer trivialen Hypothesesabbildung, die überhaupt nicht von den Merkmalen abhängt. Für numerische Labels könnte eine solche triviale Vorhersage der Stichprobenmittel der Labels in der Trainingsdatensatz sein. Bei der Klassifikation könnte die konstante Vorhersage der Label-Wert der Mehrheitsklasse in der Trainingsdatensatz sein. Ihre Genauigkeit entspricht dem Anteil dieser Klasse, der bei unausgewogenen Daten bereits hoch ist. Für den Temperatur-Datensatz aus Abb. 1 erleidet die konstante Vorhersage durch den Stichprobenmittel einen durchschnittlichen quadratischen Fehler von 87.23, fast das Sechsfache des obigen Baseline-Werts von 14.69.

Eine zweite Quelle für eine Baseline ist ein Wahrscheinlichkeitsmodell, das einen präzisen Begriff einer Baseline liefert, der auf jedes maschinelle Lernen-Verfahren anwendbar ist. Insbesondere ist das Bayes-Risiko das Risiko der optimalen Hypothese: des Bayes-Schätzers des Labels y eines Datenpunkts bei gegebenen Merkmalen $\mathbf{x}$ (Lehmann and Casella, 1998, Thm. 1.1, p. 228). Für den quadratischen Fehler ist das Bayes-Risiko gleich dem Erwartungswert der Posterior-Varianz, gebildet über die Merkmale (Bishop, 2006, Sect. 1.5.5, Eq. (1.90)). Die streifenweise Schätzung in Abb. 1 approximiert das Bayes-Risiko durch lokale Mittelung (Hastie et al., 2009, Sect. 2.4): Die Stichproben-Varianz der Labels innerhalb eines Streifens um den Merkmalswert x' approximiert die bedingte Varianz $\sigma^2_{y|x'}$. Der Durchschnitt der streifenweisen Stichproben-Varianzen approximiert den Erwartungswert $\mathbb{E}\{\sigma^2_{y|x}\}$.

Diese probabilistische Interpretation motiviert auch die obige konstante Baseline. Die konstante Vorhersage der Mehrheitsklasse ist selbst ein Bayes-Schätzer für ein Wahrscheinlichkeitsmodell $\mathbb{P}^{(y)}$, das die Merkmale der Datenpunkte ignoriert. Unter dem 0/1-Verlust weist der Bayes-Schätzer das Label mit der größten A-posteriori-Wahrscheinlichkeit zu (Bishop, 2006, Sect. 1.5.1). Für ein Wahrscheinlichkeitsmodell, das keine Merkmale berücksichtigt, ist die A-posteriori-Verteilung gleich der Prior-Verteilung des Labels, und somit ist der Bayes-Schätzer das wahrscheinlichste Label. Das Bayes-Risiko ist dann die Wahrscheinlichkeit, dass das Label einen Wert annimmt, der vom wahrscheinlichsten abweicht.

Für eine gegebene Wahl der Verlustfunktion ist der Bayes-Schätzer (sofern er existiert) vollständig durch die Wahrscheinlichkeitsverteilung $\mathbb{P}$ bestimmt (Lehmann and Casella, 1998, Thm. 1.1, p. 228). Die Berechnung des Bayes-Schätzers und des Bayes-Risikos birgt jedoch zwei wesentliche Herausforderungen. Erstens ist die Wahrscheinlichkeitsverteilung $\mathbb{P}$ unbekannt und muss aus beobachteten Daten geschätzt werden (Hastie et al., 2009, Sect. 2.4). Zweitens kann selbst dann, wenn $\mathbb{P}$ bekannt wäre, eine exakte probabilistische Inferenz für allgemeine Wahrscheinlichkeitsmodelle, wie etwa mehrfach verbundene Bayessche Netze, rechnerisch undurchführbar sein (Cooper, 1990, Sect. 4).

Ein weit verbreitetes Wahrscheinlichkeitsmodell für Datenpunkte mit numerischen Merkmalen und numerischem Label ist die multivariate Normalverteilung $(\mathbf{x}, y) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{C})$. Hier ist für den quadratischen Fehler der Bayes-Schätzer der Posterior-Mittelwert $\mu_{y|\mathbf{x}} := \mathbb{E}\{y \mid \mathbf{x}\}$ des Labels y bei gegebenen Merkmalen $\mathbf{x}$ (Gray, 2009; Lehmann and Casella, 1998, Cor. 1.2, p. 228). Das zugehörige

Bayes-Risiko ist die Posterior-Varianz $\sigma^2_{y|\mathbf{x}}$ (siehe Abb. 2). Sie ist die Baseline, mit der eine trainierte Hypothese $\widehat{h}$ verglichen wird. Abb. 2 zeigt ihre Vorhersage $\widehat{h}(\mathbf{x})$ neben dem Posterior-Mittelwert.

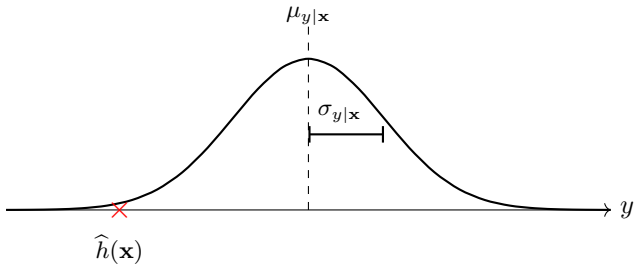

Abbildung 2: Die Wahrscheinlichkeitsverteilung (Kurve) des Labels y eines Datenpunkts, gegeben seine Merkmale $\mathbf{x}$, wenn $(\mathbf{x}, y)$ einer multivariaten Normalverteilung folgt. Die gestrichelte Linie markiert den Posterior-Mittelwert $\mu_{y|\mathbf{x}}$, den Bayes-Schätzer unter dem quadratischen Fehlerverlust; das horizontale Segment hat die Länge $\sigma_{y|\mathbf{x}}$, die Quadratwurzel der A-posteriori-Varianz $\sigma^2_{y|\mathbf{x}}$ und somit des Bayes-Risikos; das Kreuz markiert die Vorhersage $\widehat{h}(\mathbf{x})$ einer trainierten Hypothese

Für die multivariate Normalverteilung hängt diese Posterior-Varianz nicht von $\mathbf{x}$ ab (Bishop, 2006, Sect. 2.3.1, Eq. (2.73)). Genauer gesagt ist sie der Kehrwert des entsprechenden Diagonaleintrags von $\mathbf{C}^{-1}$, der Inversen der Kovarianzmatrix $\mathbf{C}$: $\sigma^2_{y|\mathbf{x}} = 1/[\mathbf{C}^{-1}]_{y,y}$. In der Praxis können der Mittelwertvektor $\boldsymbol{\mu}$ und die Kovarianzmatrix $\mathbf{C}$ aus einem Datensatz durch den Stichprobenmittel und die Stichproben-Kovarianzmatrix geschätzt werden (Anderson, 2003, Thm. 3.2.1).

Die bisherigen Baselines werden aus den Daten oder aus einem Wahrscheinlichkeitsmodell abgeleitet. Drei weitere Ansätze zur Konstruktion einer Baseline nutzen bestehende Methoden zur Bestimmung von Vorhersagen. Erstens lässt sich eine Baseline aus menschlicher Leistung gewinnen, z. B. aus der Fehlklassifikationsrate von Dermatologen, die Hautkrebs anhand von Bildern diagnostizieren (Salinas et al., 2024, Table 1). Zweitens werden für große Sprachmodelle Baselines häufig aus einem Benchmark gewonnen, d. h. aus einer festen Sammlung von Testaufgaben zusammen mit einem Auswertungsprotokoll. Beispielsweise misst der Benchmark GSM8K die Fähigkeit eines großen Sprachmodells, mathematische Textaufgaben auf Grundschulniveau zu lösen (Cobbe et al., 2021). Eine dritte Quelle ist ein bestehendes ML-Verfahren, das für die vorgesehene ML-Anwendung ungeeignet ist, z. B. weil es rechnerisch zu aufwendig ist. Der mit einem solchen Verfahren auf der Testdatensatz erzielte Fehler kann dennoch als Baseline dienen.

Siehe auch: Bayes-Risiko, Bayes-Schätzer, Testdatensatz, Wahrscheinlichkeitsmodell, Risiko, Genauigkeit, 0/1-Verlust, unausgewogene Daten, Konfusionsma-

trix, Benchmark, multivariate Normalverteilung.

Querverweise

Literatur

1. Anderson (2003). An Introduction to Multivariate Statistical Analysis. 3rd ed., Wiley, Hoboken, NJ.

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

3. Cobbe et al. (2021). Training Verifiers to Solve Math Word Problems. arXiv preprint arXiv:2110.14168. URL: https://arxiv.org/abs/2110.14168.

4. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.

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

6. Salinas et al. (2024). A systematic review and meta-analysis of artificial intelligence versus clinicians for skin cancer diagnosis. npj Digit. Med., 7(1), pp. Art. no. 125.

7. Cooper (1990). The computational complexity of probabilistic inference using Bayesian belief networks. Artif. Intell., 42(2–3), pp. 393–405.

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.