Label

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 Label ist ein Attribut eines Datenpunkt, das eine interessierende Größe darstellt und dessen Bestimmung typischerweise menschliches Fachwissen erfordert. Die Vorhersage des Labels eines Datenpunkt aus seinen Merkmale ist das Ziel des überwachtes Lernen. Die Bestimmung des Labels eines Datenpunkt ist fehleranfällig und beinhaltet oft menschliche Annotation. Die Wahl des Labels bestimmt die Lernaufgabe: Ob die Aufgabe eine Klassifikation oder eine regression ist, hängt vom Labelraum zusammen mit der Verlustfunktion ab. Beim selbstüberwachtes Lernen wird das Label aus den Merkmale selbst konstruiert, sodass keine manuelle Annotation erforderlich ist.

Definition

Für einen Datenpunkt, der einen Patienten darstellt, könnte ein Merkmal das Körpergewicht sein, während das Label das Vorhandensein einer Krankheit sein könnte. Ein Label eines Datenpunkt $\mathbf{z}$ ist eines seiner Attribute, das eine höherwertige Tatsache oder eine interessierende Größe darstellt und dessen Bestimmung typischerweise menschliches Fachwissen oder Domänenwissen erfordert (Dodge, 2003; Everitt and Skrondal, 2010; Gujarati and Porter, 2009). Im Gegensatz zu einem Merkmal, das leicht berechnet oder gemessen werden kann, ist ein Label normalerweise schwieriger zu erhalten.

Ob ein gegebenes Attribut als Merkmal oder als Label dient, ist oft eine Designentscheidung, die von den in einer gegebenen machine learning (ML)-Anwendung verfügbaren Ressourcen abhängt. Wenn ein Onkologe leicht verfügbar ist, kann eine Krebsdiagnose als Merkmal dienen. Andererseits, wenn kein Onkologe verfügbar ist, ist die Krebsdiagnose ein Label, das aus den Merkmale des Patienten vorhergesagt werden muss.

Der Begriff eines Labels wird präzisiert durch die Einführung einer Labeling-Funktion $\bar{h} : \mathcal{Z} \to \mathcal{Y}$. Der Definitionsbereich dieser Funktion ist der Datenpunkt-Raum $\mathcal{Z}$, die Menge aller möglichen Datenpunkte, die in einer gegebenen ML-Anwendung auftreten können. Seine Werte liegen im Labelraum $\mathcal{Y}$, der Menge aller möglichen Werte, die das Label eines Datenpunkt annehmen kann. Die Funktion weist jedem Datenpunkt $\mathbf{z}$ sein Label $y = \bar{h}(\mathbf{z})$ zu (Shalev-Shwartz and Ben-David, 2014, Sect. 2.1).

Die Labeling-Funktion $\bar{h}$ wirkt auf den Datenpunkt $\mathbf{z}$ selbst, das heißt, auf den gesamten Patienten. Ein trainiertes Modell (oder eine gelernte Hypothese) $\widehat{h}$ wirkt stattdessen auf den Merkmalsvektor $\mathbf{x}$, eine feste Liste von aufgezeichneten Merkmale wie dem Körpergewicht und Alter des Patienten. Der Merkmalsvektor zeichnet normalerweise nur einige Attribute eines Datenpunkt auf. Der Datenpunkt-Raum $\mathcal{Z}$ enthält Patienten, während der Merkmalsraum $\mathcal{X}$ nur deren aufgezeichnete Merkmalsvektoren enthält. Zwei Patienten mit demselben Merkmalsvektor können daher unterschiedliche Labels tragen, da sie sich in Informationen unterscheiden können, die nie als Merkmal aufgezeichnet wurden. Keine Hypothese $\widehat{h}$, die allein von $\mathbf{x}$ abhängt, kann dann perfekte Vorhersagen liefern.

In der Praxis ist das wahre Label $\bar{h}(\mathbf{z})$ eines Datenpunkt selten genau bekannt. Selbst bei vollem Zugriff auf den Datenpunkt muss dessen Label durch menschliches Urteilsvermögen oder Messung bestimmt werden, was unvollkommen ist. Das verfügbare Label ist dann eine verrauschte Version von $\bar{h}(\mathbf{z})$, wie im additiven Fehlermodell von Hastie et al. (2009), Sect. 2.6.

Zum Beispiel kann eine von einem menschlichen Annotator zugewiesene Objektkategorie oder eine von einem Kliniker erhaltene Diagnose falsch sein, obwohl der Annotator oder Kliniker den Datenpunkt direkt untersucht. Die Diskrepanz zwischen einem solchen beobachteten Label und dem wahren Label wird als Label-Rauschen bezeichnet und kann das Training beeinträchtigen und das Risiko der gelernten Hypothese erhöhen (Goodfellow et al., 2016, Sect. 7.5.1). Diese Fehlerquelle unterscheidet sich von der oben diskutierten Unvollständigkeit der aufgezeichneten Merkmale: hier ist der Datenpunkt zugänglich, aber sein Label wird unvollkommen bestimmt.

Eine Möglichkeit, den Effekt von Label-Rauschen zu reduzieren, ist Label Smoothing, das die Labels der für das Modell-Training verwendeten Datenpunkte leicht stört (Goodfellow et al., 2016, Sect. 7.5.1). Es ist eine Form der Datenerweiterung, die als Regularisierung wirkt.

Label-Rauschen ist auch für die Regulierung relevant. Der EU AI Act verlangt, dass das Trainingsdatensatz eines Hochrisiko-KI-System so weit wie möglich fehlerfrei ist (European Parliament and Council of the European Union, 2024, Art. 10(3)) und das System ein angemessenes Maß an Genauigkeit erreicht (European Parliament and Council of the European Union, 2024, Art. 15(1)); fehlerhafte Labels sind ein solcher Fehler. Wenn Labels personal data sind, verlangt die general data protection regulation (GDPR), dass sie genau sind (European Parliament and Council of the European Union, 2016, Art. 5(1)(d)).

Abgesehen davon, wie zuverlässig ein Label bestimmt wird, bestimmt welches Attribut als Label genommen wird, d.h. die Wahl der Labeling-Funktion $\bar{h}$, die Lernaufgabe. Für das Bild einer Kuhherde in Abb. 1 könnte das Label ein binärer Indikator dafür sein, ob das Bild Kühe enthält (binäre Klassifikation). Andere Optionen sind die Anzahl der Kühe oder der durchschnittliche Grünwert (regression) und ein maskiertes zentrales Pixel, das aus den umgebenden Pixeln vorhergesagt wird (selbstüberwachtes Lernen).

$y \in \{-1, 1\}$: "contains cows?" $\rightarrow$ *binäre Klassifikation*



$y \in \{1, 2, 3, \ldots\}$: "number of cows" $\rightarrow$ *regression*

$y \in \mathbb{R}$: "average green level" $\rightarrow$ *regression*

$y =$ center pixel of patch $\rightarrow$ *selbstüberwachtes Lernen*

Abbildung 1: Derselbe Datenpunkt führt je nach Wahl des Labels zu unterschiedlichen Lernaufgaben

Ob eine Lernaufgabe ein Klassifikation-Problem oder ein regression-Problem ist, hängt nicht nur vom Labelraum $\mathcal{Y}$ ab, sondern auch von der Verlustfunktion. Zum Beispiel kann der Labelraum $\mathcal{Y} = \{0, 1, 2, 3, \ldots\}$ für eine Klassifikation-Methode verwendet werden, die die 0/1 loss als Verlustfunktion nutzt. Derselbe Labelraum kann auch für eine regression-Methode verwendet werden, die die squared error loss als Verlustfunktion nutzt.

Beim selbstüberwachtes Lernen wird das Label aus den Merkmale selbst konstruiert, sodass keine manuelle Annotation erforderlich ist. Ein prominentes Beispiel ist das großes Sprachmodell (LLM)-Training: Man betrachte den Satz „Alle Menschen sind frei und gleich geboren", tokenisiert als Folge („All", „human", „beings", „are", „born", „free", „and", „equal"). Das Label für die Sub-Folge („All", „human", „beings") ist das nächste Token „are", das Label für („All", „human", „beings", „are") ist „born" und so weiter. Indem man jeden Präfix der Folge der Reihe nach nimmt, erzeugt ein einziger Satz viele gelabelte Datenpunkte allein aus einfachem Text.

Synonyme: Zielwert, Antwort, Ground Truth, interessierende Größe.

Siehe auch: Datenpunkt, Merkmal, Labelraum, Lernaufgabe, Klassifikation, regression, selbstüberwachtes Lernen, Datenerweiterung.

Querverweise

Literatur

1. Dodge (2003). The Oxford Dictionary of Statistical Terms. Oxford Univ. Press, New York, NY, USA.

2. European Parliament and Council of the European Union (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and

(EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). Official Journal of the European Union, L series, Jul. 12. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.

3. Everitt and Skrondal (2010). The Cambridge Dictionary of Statistics. 4th ed., Cambridge Univ. Press, Cambridge, U.K.

4. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

5. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

6. Gujarati and Porter (2009). Basic Econometrics. 5th ed., McGraw-Hill/Irwin, New York, NY, USA.

7. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

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.