Sobolev-Raum

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 Sobolev-Raum ist eine Menge von Funktionen, deren Ableitungen in einem verallgemeinerten Sinn eine beschränkte Größe haben. Methoden des maschinelles Lernen (ML), deren Hypothesenraum eine Kugel in einem Sobolev-Raum mit einer Schranke für den Gradienten ist, erzwingen einen wohldefinierten quantitativen Begriff von Robustheit. Jede Hypothesenabbildung in einer solchen Kugel lässt sich lokal durch eine lineare Abbildung mit beschränkter Steigung approximieren. Eine kleine Änderung der Merkmale eines Datenpunkts ändert daher die Vorhersage höchstens um einen proportionalen Betrag. Die verallgemeinerte Ableitung wird benötigt, da Hypothesenabbildungen wie die eines tiefes Netz mit ReLU (ReLU)-Aktivierungsfunktionen Knicke aufweisen. Ein Entscheidungsbaum, dessen Hypothesenabbildung springt, gehört keinem Sobolev-Raum an und lässt keine solche Garantie zu. Eine lineare Regression, deren Hypothesenraum eine Kugel in einem Sobolev-Raum anstelle der lineare Abbildungen ist, stellt eine nichtparametrische Regression dar. Sobolev-Normen dienen auch als Regularisierungsterme und beschreiben den RKHS (RKHS) einer Matérn-Kernel.

Definition

Ein Wetterdienst sagt die Lufttemperatur um 15:00 Uhr (UTC) eines Tages aus dem Messwert von 06:00 Uhr (UTC) voraus. Abb. 1(a) zeigt 20 der 365 Tage des Jahres 2024 an der Station Krems an der Donau der GeoSphere Austria als die Datenpunkte eines Streudiagramms, mit dem Messwert eines Tages um 06:00 Uhr als seinem Merkmal und dem Messwert um 15:00 Uhr als seinem Label. Die Teilabbildung zeigt außerdem drei aus diesen Datenpunkten gelernte Hypothesen: die nichtlineare Abbildung eines tiefes Netz mit ReLU (ReLU)-Aktivierungsfunktionen, die TreppenFunktion eines Entscheidungsbaums der Tiefe 2 und die Kurve einer lineare Regression mit einem Sobolev-Strafterm. Ein vertrauenswürdiges Vorhersagemodell darf seine Vorhersage, die Vorhersage, nicht sprunghaft ändern: Ein Thermometerfehler von einem halben Grad beim Messwert um 06:00 Uhr darf die Vorhersage nicht um mehrere Grad verschieben. Sobolev-Räume präzisieren diese Anforderung. Sie bestehen aus den Funktionen, deren Ableitungen eine beschränkte Größe haben (Evans, 2010,

Ch. 5). Eine Schranke für die Ableitung ist eine Schranke dafür, wie schnell sich die Vorhersage ändern kann.

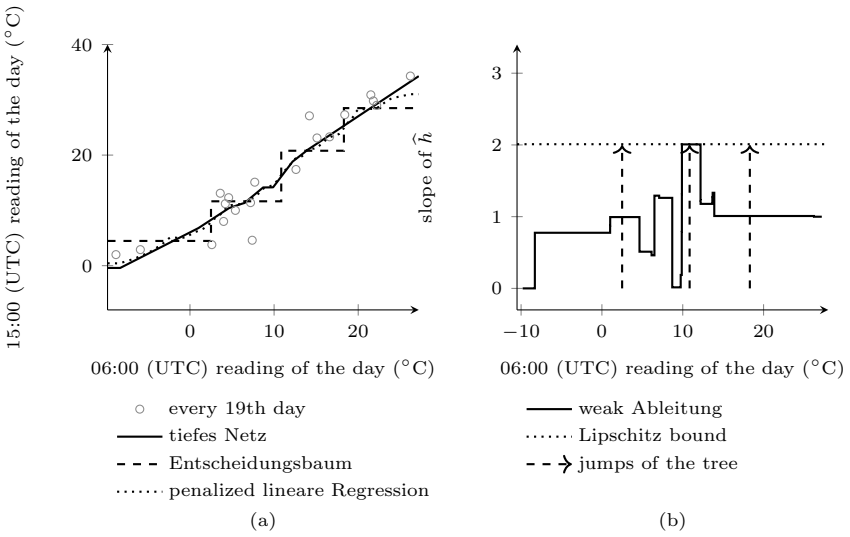


Abbildung 1: Vorhersage der Lufttemperatur um 15:00 (UTC) eines Tages in Krems an der Donau aus dem Messwert um 06:00 (UTC). (a) Streudiagramm jedes 19. Tages im Jahr 2024, 20 der 365 Tage (Kreise), sowie drei aus allen 365 Tagen gelernte Hypothesen: die nichtlineare Abbildung eines tiefes Netz mit ReLU-Aktivierungsfunktionen (durchgezogene Kurve), die Treppen-Funktion eines Entscheidungsbaums der Tiefe 2 (gestrichelte Kurve) und die Kurve einer lineare Regression mit einem Sobolev-Strafterm (gepunktete Kurve). (b) Die Steigung der tiefes Netz-Hypothese $\widehat{h}$ ist eine Treppen-Funktion, ihre schwache Ableitung, und bleibt unter der Schranke $L = 2.01$ (gepunktete Linie), sodass eine Störung δ des Messwerts um 06:00 die Vorhersage um höchstens $2.01\,|\delta|$ verschiebt. Der Baum besitzt keine solche Schranke: An jedem seiner Schwellenwerte (Pfeile) springt die Vorhersage. Daten erzeugt mit `pythondemos/sobolevspace.py`

Die Ableitung wird in einem verallgemeinerten Sinn verstanden, da das tiefes Netz aus Abb. 1(a) Knicke aufweist und dort keine klassische Ableitung existiert. Die verallgemeinerte Ableitung geht von der Integration statt von der Differentiation aus. Auf der reellen Achse ist eine Funktion mit $f(x) = f(a) + \int_a^x g(t)\,\mathrm{d}t$ das Integral ihrer Steigung g, und g wird als schwache Ableitung von f bezeichnet. Auf einer offenen Menge $\Omega \subseteq \mathbb{R}^d$ betrachte man eine unendlich oft differenzierbare TestFunktion φ, die außerhalb einer beschränkten Teilmenge von Ω verschwindet. Eine Funktion g_j ist eine schwache partielle Ableitung von

f, wenn für jedes solche φ gilt (Evans, 2010, Sect. 5.2):

$$\int_\Omega f(\mathbf{x}) \frac{\partial \varphi(\mathbf{x})}{\partial x_j} \,\mathrm{d}\mathbf{x} = -\int_\Omega g_j(\mathbf{x})\varphi(\mathbf{x}) \,\mathrm{d}\mathbf{x}.$$

Dies ist partielle Integration, bei der alle Ableitungen auf φ verlagert sind. Sie gilt, wenn f das Integral von g_j entlang der j-ten Koordinate ist, und setzt in keinem einzelnen Punkt eine Ableitung von f voraus. Die Steigung des tiefes Netz, die TreppenFunktion aus Abb. 1(b), erfüllt diese Gleichung für jede geprüfte TestFunktion (`pythondemos/sobolevspace.py`).

Für $p \in [1, \infty]$ besteht der Sobolev-Raum $W^{1,p}(\Omega)$ aus den Funktionen, deren schwache Ableitungen existieren und eine endliche L^p-Norm $\|\cdot\|_p$ haben. Der Fall $p = 2$ liefert den Hilbertraum $H^1(\Omega)$, und die Räume $W^{k,p}(\Omega)$ verlangen schwache Ableitungen bis zur Ordnung k (Adams and Fournier, 2003, Ch. 3, Thm. 3.6).

Der für das maschinelles Lernen (ML) wichtigste Fall ist $W^{1,\infty}$, die Menge der Funktionen mit einem beschränkten schwachen Gradienten. Auf einer konvexen Menge Ω gehört eine Funktion f mit $\operatorname{ess\,sup}_{\mathbf{x}\in\Omega} \|\nabla f(\mathbf{x})\|_2 \leq L$ genau dann dazu, wenn sie Lipschitz-Stetigkeit-stetig mit der Konstanten L ist (Evans, 2010, Sect. 5.8). Eine Störung eines Merkmalsvektors um $\boldsymbol{\delta}$ ändert f dann um höchstens $L\|\boldsymbol{\delta}\|_2$: Die Konvexität hält die Verbindungsstrecke von $\mathbf{x}$ nach $\mathbf{x} + \boldsymbol{\delta}$ innerhalb von Ω, und die Integration des Gradienten entlang dieser Strecke beschränkt die Änderung durch $L\|\boldsymbol{\delta}\|_2$. Umgekehrt existiert der Gradient einer Lipschitz-Stetigkeit-stetigen Funktion f fast überall, und wo er existiert, ist seine Länge die Steigung von f in seiner eigenen Richtung, also höchstens L. Das tiefes Netz aus Abb. 1 hat $L = 2.01$, sodass ein Fehler von 0.5 °C beim Messwert um 06:00 Uhr seine Vorhersage um höchstens 1.00 °C verschiebt; die größte Änderung unter 20000 zufälligen Störungen dieser Größe betrug 1.00 °C. Die Garantie stammt aus der Schranke, nicht aus dem Raum: $W^{1,\infty}$ enthält Funktionen mit jedem Wert von L. Eine Methode des ML, deren Hypothesenraum die Kugel von Funktionen mit einer Steigung von höchstens L ist, lernt konstruktionsbedingt eine robuste Hypothese, unabhängig vom Trainingsdatensatz. Die Wahl $p = \infty$ bewirkt, dass die Schranke bei jedem Merkmalsvektor gilt; H^1 beschränkt die Steigung nur im Mittel. Die Schranke betrifft ausschließlich Störungen der Merkmale; Robustheit gegenüber Störungen des Trainingsdatensatzs ist eine Eigenschaft der Trainingsabbildung (siehe Stabilität).

Die Schranke besagt außerdem, dass die Hypothese in der Nähe fast jedes Merkmalsvektors mit einer lineare Abbildung übereinstimmt, deren Steigung höchstens L beträgt (Satz von Rademacher, Evans, 2010, Sect. 5.8.3, Thm. 6). Diese lineare Abbildung ist die Erklärung, die gradientenbasierte erklärbare künstliche Intelligenz (XAI) liefert und die lokale interpretierbare modellunabhängige Erklärungen (LIME) anhand gestörter Datenpunkte anpasst. Für die Vorhersage lautet sie: Ein zusätzliches Grad beim Messwert um 06:00 Uhr fügt der Vorhersage $s(x)$ Grad hinzu, wobei die Steigung $s(x)$ aus Abb. 1(b) zwischen 0 und 2.01 liegt. Da das tiefes Netz stückweise linear ist, ist die Approximation exakt, sobald für eine Störung δ des Messwerts um 06:00 Uhr kein

Knick zwischen x und $x+\delta$ liegt, was bei 98 Prozent der Paare mit $|\delta| \leq 0.1\,°\mathrm{C}$ der Fall ist.

Für das tiefes Netz ist der schwache Gradient durch das Produkt der Spektralnormen seiner Gewichtsmatrizen beschränkt, d. h. ihrer größten Singulärwerte (Tsuzuku et al., 2018, Sects. 5.1 and 5.2), hier 38.0 gegenüber dem wahren Wert 2.01: gültig, aber grob. Für einen Klassifikator, der das Vorzeichen eines solchen f verwendet, zertifiziert die Schranke, dass keine Störung mit einer euklidische Norm unterhalb von $|f(\mathbf{x})|/L$ die Vorhersage ändert, da $|f(\mathbf{x}+\boldsymbol{\delta}) - f(\mathbf{x})| \leq L\|\boldsymbol{\delta}\|_2 < |f(\mathbf{x})|$ das Vorzeichen von f unverändert lässt (Tsuzuku et al., 2018, Prop. 2). Der Entscheidungsbaum aus Abb. 1(a) besitzt keine solche Schranke: An seinem Wurzelschwellenwert von 10.85 °C verschiebt eine Änderung von $2 \cdot 10^{-9}\,°\mathrm{C}$ beim Messwert um 06:00 Uhr die Vorhersage um 9.14 °C. Ein Sprung besitzt keine schwache Ableitung, die eine Funktion ist (Evans, 2010, Sect. 5.2, Example 2), sodass kein Sobolev-Raum den Baum enthält.

Sobolev-Räume dienen auch eigenständig als Hypothesenräume. Die gewöhnliche lineare Regression minimiert den mittleren quadratischer Verlust über die lineare Abbildungen. Ersetzt man die lineare Abbildungen durch eine Kugel in H^1 oder fügt man äquivalent die quadrierte H^1-Halbnorm $\int_\Omega \|\nabla f(\mathbf{x})\|_2^2 \,\mathrm{d}\mathbf{x}$ als Strafterm hinzu, eine Norm, die konstante Verschiebungen von f ignoriert (siehe Regularisierung), so erhält man eine nichtparametrische Regression (Wainwright, 2019, Sects. 13.1.2 and 13.4). Die gepunktete Kurve in Abb. 1(a) minimiert den Trainingsfehler zuzüglich des Zehnfachen dieser Halbnorm; sie erreicht einen Trainingsfehler von 13.92, unterhalb der 14.56 des tiefes Netz, bei einer Steigung von höchstens 2.82 (`pythondemos/sobolevspace.py`).

Ein Computer kann nicht ganz H^1 durchsuchen, und das muss er auch nicht. Für ein skalares Merkmal, bei dem Ω ein Intervall der reellen Achse ist, ist ein Minimierer aus Trainingsfehler plus Halbnorm stückweise linear mit Knicken an den m Datenpunkten, da zwischen zwei Datenpunkten die Gerade bei gegebenen Endwerten die kleinste Halbnorm aufweist (eine Folgerung aus der Cauchy-Schwarz-Ungleichung). Die Suche läuft daher über m Zahlen, die Werte an den Knicken; dies ist die Kernel-Ridge-Regression mit dem RKHS (RKHS) von H^1 (Wainwright, 2019, Sect. 12.5.2 and Example 12.16). Für eine stückweise lineare Funktion ist die Halbnorm eine Summe quadrierter Steigungen, eine pro Abschnitt und gewichtet mit dessen Länge, sodass die Normalgleichungen des bestraften Problems ein einziges lineares Gleichungssystem bilden.

Die Halbnorm bringt die Glattheitsannahme zum Ausdruck. Ihr Gegenstück für eine Funktion auf den Knoten eines Graphen ersetzt den Gradienten durch Differenzen entlang der Kanten, jene Form, die von der verallgemeinerte totale Variation (GTV) und der verallgemeinerte Totalvariationsminimierung (GTVMin) verwendet wird (SarcheshmehPour et al., 2023); bestraft man den Gradienten stattdessen in L^1, ergibt sich die Totalvariation-Regularisierung, die Sprünge zulässt (Chambolle and Pock, 2016, Sect. 7.1; Rudin et al., 1992). Der RKHS einer Matérn-Kernel ist ein Sobolev-Raum $W^{s,2}(\Omega)$, dessen Ordnung s durch den Glattheits-Parameter der Kernel festgelegt wird (Kanagawa et al., 2018, Example 2.6). Eine Kernel-Methode mit dieser Kernel ist daher wiederum eine

nichtparametrische Regression über einer Sobolev-Kugel, der oben beschriebene Rahmen aus Wainwright (2019), Ch. 13.

Schließlich ist eine Schranke L auch von regulatorischem Interesse. Artikel 15 des KI-Verordnung verlangt, dass ein Hochrisiko-KI-System robust ist und seine Metriken zur Genauigkeit in der Gebrauchsanweisung angegeben werden (European Parliament and Council of the European Union, 2024). Eine Lipschitz-Stetigkeit-Konstante ist eine solche Metrik: Sie gibt als einzelne Zahl an, wie stark eine Störung eines Merkmalsvektors die Vorhersage verändern kann.

Siehe auch: Lipschitz-Stetigkeit, Robustheit, Stabilität, tiefes Netz, Entscheidungsbaum, lineare Regression, Regularisierung, XAI, LIME, Glattheitsannahme, GTV, GTVMin, RKHS, Hilbertraum, differenzierbar.

Querverweise

Literatur

1. Adams and Fournier (2003). Sobolev Spaces. 2nd ed., Academic Press, Amsterdam, The Netherlands.

2. SarcheshmehPour et al. (2023). Clustered Federated Learning via Generalized Total Variation Minimization. IEEE Trans. Signal Process., 71, pp. 4240–4256.

3. 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.

4. Wainwright (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press, Cambridge, UK.

5. Evans (2010). Partial Differential Equations. 2nd ed., American Mathematical Society, Providence, RI, USA.

6. Kanagawa et al. (2018). Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. URL: https://arxiv.org/abs/1807.02582.

7. Chambolle and Pock (2016). An introduction to continuous optimization for imaging. Acta Numer., 25, pp. 161–319.

8. Rudin et al. (1992). Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1–4), pp. 259–268.

9. Tsuzuku et al. (2018). Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 6541–6550.