effektive Dimension

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 effektive Dimension eines Hypothesenraums ist seine durch die Wahrscheinlichkeitsverteilung der Merkmale gesehene Größe: die Anzahl der Modellparameter, die auf den von der Verteilung erzeugten Datenpunkten einen Unterschied machen. Für ein lineares Modell ist sie der Rang der Matrix der zweiten Momente der Merkmalsabbildung, geschätzt durch den Rang der Merkmalsmatrix, und sie übersteigt nie die Anzahl der Modellparameter. Zwei Hypothesenräume, die sich nur außerhalb des Trägers der Verteilung unterscheiden, haben dieselbe effektive Dimension, und die empirische Risikominimierung auf ihnen liefert auf den Daten dieselben Vorhersagen. Derselbe Hypothesenraum kann unter verschiedenen Datenverteilungen verschiedene effektive Dimensionen haben. Die Rademacher-Komplexität ist das verteilungsabhängige Größenmaß der Lerntheorie, während die Vapnik-Chervonenkis-Dimension (VC-Dimension) ihr ungünstigster Fall über den gesamten Merkmalsraum ist.

Definition

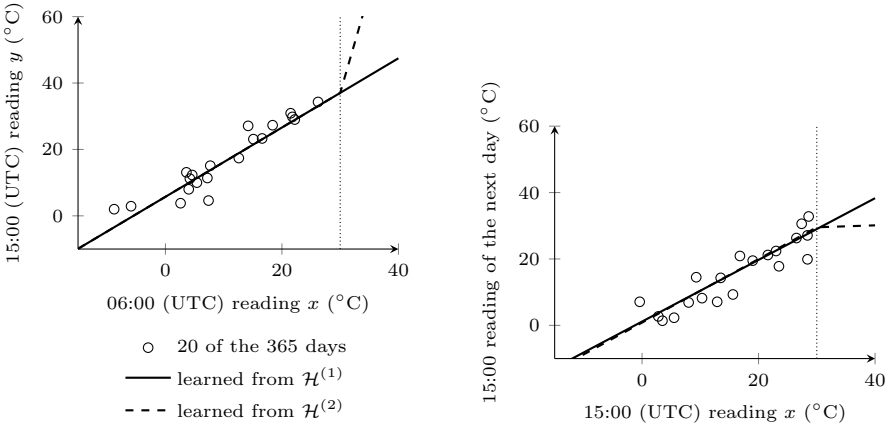
Betrachtet seien die 365 Tage des Jahres 2024 an der Wetterstation Krems an der Donau (Österreich), für die beide Messwerte vorliegen, jeder ein Datenpunkt mit der Lufttemperatur um 06:00 Uhr (UTC) als seinem Merkmal x und dem Messwert von 15:00 Uhr (UTC) als seinem Label y. Zwei Hypothesenräume stehen für die Vorhersage von y aus x zur Auswahl: das affine lineare Modell $\mathcal{H}^{(1)} = \{w_0 + w_1 x\}$ mit zwei Modellparametern und das lineare Modell $\mathcal{H}^{(2)} = \{w_0 + w_1 x + w_2 \max(0, x - 30)\}$ mit dreien, dessen drittes Merkmal ein ReLU (ReLU)-Knickpunkt bei $30\,^\circ$C ist. Der 06:00-Uhr-Messwert der Aufzeichnung reicht von -8.8 bis $26.2\,^\circ$C, sodass der Knickpunkt an jedem Tag null ist. Auf dieser Aufzeichnung stimmt jede Hypothese aus $\mathcal{H}^{(2)}$ mit einer Hypothese aus $\mathcal{H}^{(1)}$ überein: Der dritte Modellparameter ändert keine Vorhersage, die anhand der Daten überprüft werden kann.

Die effektive Dimension $d_{\text{eff}}(\mathcal{H})$ eines Hypothesenraums $\mathcal{H}$ ist seine durch die Wahrscheinlichkeitsverteilung $\mathbb{P}^{(x)}$ der Merkmale gesehene Größe: Sie zählt die Modellparameter, die auf den von der Verteilung erzeugten Datenpunkten einen Unterschied machen, nicht die Modellparameter, die hingeschrieben werden. Für ein lineares Modell mit Merkmalsabbildung $\boldsymbol{\Phi}$ ist die effektive Dimension definiert als der Rang der Matrix der zweiten Momente $\mathbb{E}\{\boldsymbol{\Phi}(x)\boldsymbol{\Phi}(x)^\top\}$. Ein

Datensatz schätzt sie durch den Rang seiner Merkmalsmatrix, was der Spur der orthogonalen Projektion entspricht, welche die kleinste Quadrate auf den Labelvektor anwendet – der effektiven Anzahl von Modellparametern eines Verfahrens, dessen Vorhersagen eine lineare Funktion der Labels sind (Hastie et al., 2009, Abschn. 7.6). Sie übersteigt nie die Anzahl der Modellparameter und gleicht dieser Anzahl nur dann, wenn die Merkmale auf dem Träger von $\mathbb{P}^{(x)}$ linear unabhängig sind. Auf der Kremser Aufzeichnung hat die Merkmalsmatrix von $\mathcal{H}^{(2)}$ drei Spalten und den Rang 2, da ihre Knickpunktspalte null ist, sodass $d_{\text{eff}}\left(\mathcal{H}^{(1)}\right) = d_{\text{eff}}\left(\mathcal{H}^{(2)}\right) = 2$ gilt. Zwei Hypothesenräume, die sich nur außerhalb des Trägers von $\mathbb{P}^{(x)}$ unterscheiden, haben dieselbe effektive Dimension.

Die Konsequenz für die empirische Risikominimierung ist in Abb. 1(a) zu sehen. Die Methode der kleinsten Quadrate auf $\mathcal{H}^{(1)}$ und auf $\mathcal{H}^{(2)}$ liefert dieselbe Vorhersage für jeden Tag der Aufzeichnung und denselben Trainingsfehler von 15.25 Quadratgrad. Der Knickpunktkoeffizient w_2 ist nicht identifizierbar: Jeder Wert ergibt denselben Trainingsfehler, und die beiden gelernten Hypothesen unterscheiden sich nur für einen 06:00-Uhr-Messwert oberhalb von 30 °C, einem Bereich, den kein Tag der Aufzeichnung erreicht. Dort sagen die Daten nichts aus, was das Problem der Generalisierung in seiner schärfsten Form darstellt: Eine Hypothese wird nur dort geprüft, wo die Merkmale liegen.

Die effektive Dimension hängt von der Wahrscheinlichkeitsverteilung ab, nicht nur vom Hypothesenraum. Abb. 1(b) behält $\mathcal{H}^{(2)}$ bei und ändert das Merkmal auf den 15:00-Uhr-Messwert des Tages, mit dem 15:00-Uhr-Messwert des nächsten Tages als Label. Der 15:00-Uhr-Messwert überschreitet 30 °C an 30 der 363 Tagespaare, welche die Aufzeichnung an zwei aufeinanderfolgenden Tagen enthält; die Knickpunktspalte der Merkmalsmatrix ist nicht mehr null, ihr Rang beträgt 3, und die Methode der kleinsten Quadrate erreicht auf $\mathcal{H}^{(2)}$ nun einen kleineren Trainingsfehler als auf $\mathcal{H}^{(1)}$, wenn auch nur geringfügig. Derselbe Hypothesenraum hat unter der einen Datenverteilung die effektive Dimension 2 und unter einer anderen 3.



(a) the 06:00 reading as the Merkmal

(b) the 15:00 reading as the Merkmal

Abbildung 1: Dieselben zwei Hypothesenräume unter zwei Datenverteilungen in Krems an der Donau, 2024. In (a) ist das Merkmal der 06:00-Uhr-Messwert (UTC) des Tages, der den Knickpunkt bei 30 °C (gepunktete vertikale Linie) nie erreicht: Die mittels kleinster Quadrate aus $\mathcal{H}^{(1)}$ (durchgezogene Linie) und aus $\mathcal{H}^{(2)}$ (gestrichelte Linie, gezeichnet mit dem auf 5 gesetzten Knickpunktkoeffizienten) gelernten Hypothesen fallen überall dort zusammen, wo die Daten liegen, und unterscheiden sich nur jenseits davon. In (b) ist das Merkmal der 15:00-Uhr-Messwert des Tages und das Label der 15:00-Uhr-Messwert des nächsten Tages; der Knickpunkt ist an den heißen Tagen aktiv, und die beiden gelernten Hypothesen unterscheiden sich geringfügig jenseits von 30 °C. Offene Kreise stellen 20 der Tage dar, jeden 19 in zeitlicher Reihenfolge. Wetterdaten: GeoSphere Austria (GeoSphere Austria, 2026). Daten generiert durch `pythondemos/effdim.py`

Über lineare Modelle hinaus unterscheiden sich die Größenmaße der Lerntheorie auch darin, ob sie einen ungünstigsten Fall über den Merkmalsraum oder einen Durchschnitt über die Datenverteilung betrachten. Die Vapnik-Chervonenkis-Dimension (VC-Dimension) ist ein ungünstigster Fall über alle Datensätze, die der Merkmalsraum zulässt (Shalev-Shwartz and Ben-David, 2014, Def. 6.5), sodass sie den Knickpunkt von $\mathcal{H}^{(2)}$ mitzählt, unabhängig davon, ob die Daten ihn jemals erreichen oder nicht: Für die durch Schwellenwertbildung einer Hypothese bei null erhaltenen Klassifikatoren werden drei Merkmal-Werte, von denen einer jenseits des Knickpunkts liegt, von $\mathcal{H}^{(2)}$ zerschmettert, da die dortigen drei Werte einer Hypothese ein invertierbares lineares Bild von (w_0, w_1, w_2) sind, während jede Hypothese aus $\mathcal{H}^{(1)}$ monoton in x ist und somit keine Markierung realisiert, die zweimal das Vorzeichen wechselt. Die Rademacher-Komplexität ist ein Erwartungswert über aus $\mathbb{P}^{(x)}$ gezogene Datensätze und daher die effektive Größe von $\mathcal{H}$ für die vorliegende Anwendung (Bartlett and Mendelson, 2002, Def. 2). Für ein Verfahren wie die Ridge-Regression, dessen gelernte Hypothese eine lineare Funktion der Labels

ist, ist die Spur dieser linearen Abbildung die effektive Anzahl der Modellparameter: Sie nimmt mit wachsendem Regularisierungsparameter α vom Rang der Merkmalsmatrix gegen null ab (Hastie et al., 2009, Abschn. 7.6).

Siehe auch: Hypothesenraum, Modellparameter, Dimension, lineares Modell, Merkmalsabbildung, Merkmalsmatrix, Rang, Wahrscheinlichkeitsverteilung, Generalisierung, VC-Dimension, Rademacher-Komplexität, Ridge-Regression.

Querverweise

Literatur

1. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Data hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/hist v2-1d.

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

3. Bartlett and Mendelson (2002). Rademacher and gaussian complexities: Risk bounds and structural results. J. Mach. Learn. Res., 3(Nov), pp. 463–482.

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