Hypothesenraum

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 Hypotheseraum $\mathcal{H}$ ist eine Menge von Hypotheseabbildungen $h : \mathcal{X} \to \mathcal{Y}$ von einem Merkmalsraum in einen Labelraum. Jede machine learning (ML)-Methode verwendet einen zugrunde liegenden Hypotheseraum, der eine Teilmenge der Menge $\mathcal{Y}^{\mathcal{X}}$ aller möglichen Abbildungen von $\mathcal{X}$ nach $\mathcal{Y}$ ist. Die verfügbaren Rechenressourcen begrenzen die Größe von $\mathcal{H}$, was wiederum die Rechenkosten der Methode und die generalization gap bestimmt. Der Hypotheseraum trägt typischerweise eine geometrische Struktur: Reellwertige Hypothesen können über die Abweichung ihrer Vorhersagen auf einem Referenz- Datensatz verglichen werden, und ein parametrisches Modell erbt eine geometrische Struktur vom zugrunde liegenden parameter space $\mathcal{W}$.

Definition

Ein Wetterdienst, der die morgige Temperatur um 15:00 (UTC) aus dem heutigen Messwert um 06:00 vorhersagt, muss sich auf eine Menge von Kandidaten-Vorhersageabbildungen festlegen: zum Beispiel alle linearen Funktionen der Morgentemperatur. Diese Kandidatenmenge ist der Hypothesenraum der Methode; allgemeiner legt sich jede machine learning (ML)-Methode auf einen bestimmten Hypothesenraum $\mathcal{H}$ fest, aus dem sie eine einzelne Hypothese $\widehat{h}$ lernt. Formal ist $\mathcal{H}$ eine Teilmenge der Menge $\mathcal{Y}^{\mathcal{X}}$ aller möglichen Abbildungen vom Merkmalsraum in den Labelraum (Shalev-Shwartz and Ben-David, 2014, Sekt. 2.3; Vapnik, 2000, Sekt. 1.1) (siehe Abb. 1).

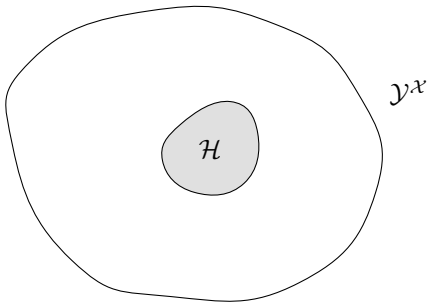

Abbildung 1: Hypothesenraum $\mathcal{H}$ einer ML-Methode als eingeschränkte Teilmenge der viel größeren Menge $\mathcal{Y}^{\mathcal{X}}$ aller möglichen Abbildungen vom Merkmalsraum $\mathcal{X}$ in den Labelraum $\mathcal{Y}$

Lineare Regression und andere lineare Methoden verwenden das lineares Modell als $\mathcal{H}$, also die Menge aller lineare Abbildungen $\mathbb{R}^d \to \mathbb{R}$ (Shalev-Shwartz and Ben-David, 2014, Sekt. 9.2). Ein weiteres kanonisches Beispiel ist die Menge der durch ein künstliches neuronales Netz (KNN) mit fester Architektur realisierbaren Eingabe-Ausgabe-Abbildungen, wenn die Modellparameter variieren.

Der gleiche $\mathcal{H}$ kann von verschiedenen ML-Algorithmen verwendet werden, etwa von empirische Risikominimierung (ERM) auf einem festen Trainingsdatensatz, von Online-Lernen auf einem Strom von Datenpunkte oder von Bayesian inference, die eine A-posteriori-Verteilung über $\mathcal{H}$ statt einer einzelnen Hypothese zurückgibt (Bishop, 2006, Sekt. 1.2.3).

Die Wahl von $\mathcal{H}$ ist eine zentrale Entwurfsentscheidung für eine ML-Methode. Aus Sicht des ML-Engineerings wird sie von zwei Faktoren bestimmt. Der erste sind die verfügbaren Rechenressourcen wie Speicher, Rechenzeit und Kommunikationsbandbreite. Der zweite Faktor ist die Anzahl verfügbarer Datenpunkte, die die maximal effektiv trainierbare Größe von $\mathcal{H}$ ohne Überanpassung begrenzt. Prinzipiell könnte $\mathcal{H}$ so groß wie $\mathcal{Y}^{\mathcal{X}}$ selbst sein. Die Größe von $\mathcal{H}$ erlaubt eine Abschätzung der Generalisierung, noch bevor ein Training stattfindet. Ein kleiner $\mathcal{H}$ führt typischerweise zu einer kleinen generalization gap, während ein großer $\mathcal{H}$ anfälliger für Überanpassung ist (Shalev-Shwartz and Ben-David, 2014, Kor. 4.6, Thm. 6.8). Für einen endlichen $\mathcal{H}$ ist die effektive Größe seine Mächtigkeit. Für einen unendlichen $\mathcal{H}$ kann die effektive Größe durch die Vapnik–Chervonenkis dimension (VC dimension) oder die Rademacher complexity quantifiziert werden (Shalev-Shwartz and Ben-David, 2014, Thm. 6.8, Thm. 26.5).

Das Wort *Raum* (statt *Menge*) spiegelt wider, dass $\mathcal{H}$ typischerweise eine geometrische Struktur trägt. Im Allgemeinen ist $\mathcal{H}$ kein Vektorraum: Die Summe zweier Hypothesen in $\mathcal{H}$ muss nicht zu $\mathcal{H}$ gehören. Darüber hinaus gibt es Hypothesenräume, die überhaupt keine algebraische Struktur besitzen, d. h., es gibt keine sinnvolle Möglichkeit, zwei Hypothesen in $\mathcal{H}$ zu addieren oder eine Hypothese mit einem Skalar zu multiplizieren.

Für reellwertige Hypothesen können zwei $h, h' \in \mathcal{H}$ über ihre mittlere quadratische Vorhersageabweichung auf einem Referenz-Datensatz $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\}$ verglichen werden,

$$\frac{1}{m} \sum_{r=1}^{m} \left(h(\mathbf{x}^{(r)}) - h'(\mathbf{x}^{(r)})\right)^2.$$

Abb. 2 illustriert diesen Vergleich: Die beiden Hypothesen werden an den Merkmal-Werten der Datenpunkte in $\mathcal{D}$ ausgewertet und die quadrierten Differenzen ihrer Vorhersagen gemittelt. Die Quadratwurzel dieses Mittels, $d_{\mathcal{D}}(h, h')$, ist eine Pseudo-Metrik auf $\mathcal{H}$: Sie ist die Euclidean norm der Differenz der beiden Vorhersage-Vektoren geteilt durch $\sqrt{m}$, also symmetrisch und erfüllt die Dreiecksungleichung, und sie verschwindet für zwei verschiedene Hypothesen, die auf $\mathcal{D}$ übereinstimmen. Diese Konstruktion ist für Hypothesen mit einer endlichen Labelraum, wie sie in Klassifikation-Methoden verwendet werden, nicht sinnvoll. Stattdessen können zwei solche Hypothesen über ihre Unstimmigkeitsrate auf $\mathcal{D}$ verglichen werden.

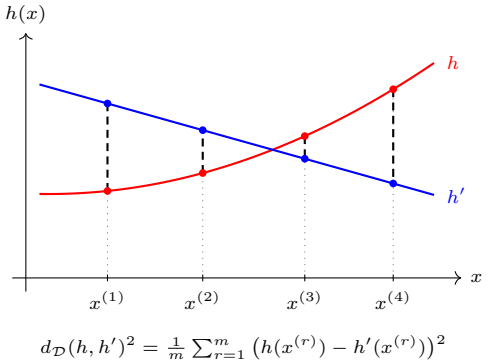

Abbildung 2: Vergleich zweier Hypothesen h (rot) und h' (blau) anhand ihrer Vorhersagen bei vier skalaren Merkmale $x^{(1)}, \ldots, x^{(4)}$. Die gestrichelten Strecken markieren die Differenz $h(x^{(r)}) - h'(x^{(r)})$ für jeden Datenpunkt; die Quadratwurzel des Mittels ihrer Quadrate ist die Pseudo-Metrik $d_{\mathcal{D}}(h, h')$ zwischen h und h'

Ein parametrisches Modell $\mathcal{H} = \{h^{(\mathbf{w})} : \mathbf{w} \in \mathcal{W}\}$ erbt eine geometrische Struktur vom zugrunde liegenden parameter space $\mathcal{W} \subseteq \mathbb{R}^d$. Eine Norm auf $\mathbb{R}^d$ induziert die Kandidaten-Abstand $d(h^{(\mathbf{w})}, h^{(\mathbf{w}')}) := \|\mathbf{w} - \mathbf{w}'\|$ auf $\mathcal{H}$. Ist die Parametrisierung $\mathbf{w} \mapsto h^{(\mathbf{w})}$ injective, so liefern verschiedene Gewicht-Vektoren verschiedene Hypothesen, und diese Kandidaten-Abstand ist eine Metrik. Das lineares Modell $\mathcal{H} = \{h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} : \mathbf{w} \in \mathbb{R}^d\}$ ist ein kanonisches Beispiel für eine injective Parametrisierung: Verschiedene Gewicht-Vektoren $\mathbf{w}$ erzeugen verschiedene lineare Abbildungen, sodass $\|\mathbf{w} - \mathbf{w}'\|$ eine Metrik auf $\mathcal{H}$ ist. Ist die Parametrisierung nicht injective, so können zwei verschiedene Gewicht-Vektoren $\mathbf{w} \neq \mathbf{w}'$ die gleiche Hypothese $h^{(\mathbf{w})} = h^{(\mathbf{w}')}$ repräsentieren.

Die Abstand $\|\mathbf{w} - \mathbf{w}'\|$ ist dann strikt positiv, obwohl die beiden Hypothesen übereinstimmen, also nur eine Pseudo-Metrik. Das Vertauschen der versteckten Einheiten eines KNN etwa ändert $\mathbf{w}$, lässt aber $h^{(\mathbf{w})}$ unverändert (Bishop, 2006, Sekt. 5.1.1).

Synonyme: Hypothesenklasse, Modell.

Siehe auch: Hypothese, Modell, lineares Modell, parametrisches Modell, parameter space, VC dimension, Rademacher complexity, ERM.

Querverweise

Literatur

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

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

3. Vapnik (2000). The Nature of Statistical Learning Theory. 2nd ed., Springer, New York, NY, USA.