Generalisierung

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 Generalisierung ist die Fähigkeit einer machine learning (ML)-Methode, genaue Vorhersagen auf Datenpunkte zu treffen, die während des Training nicht verwendet wurden. Viele ML-Methoden implementieren das Training über empirische Risikominimierung (ERM), d.h. das Lernen einer Hypothese $\widehat{h}$, die das empirical risk auf dem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ minimiert. Die gelernte $\widehat{h}$ wird dann verwendet, um Vorhersagen auf neuen Datenpunkte außerhalb von $\mathcal{D}^{(\text{train})}$ zu berechnen. Ein Wahrscheinlichkeitsmodell zur Daten-Generierung quantifiziert die Vorhersage-Fehler jenseits von $\mathcal{D}^{(\text{train})}$ über das Risiko der $\widehat{h}$, d.h. den erwarteten Verlust, der bei einem neuen zufällig ausgewählten Datenpunkt entsteht. Die Differenz zwischen dem Risiko und dem empirical risk ist die generalization gap. Die stability der ML-Methode begrenzt die erwartete generalization gap ohne Bezug auf die Größe des Hypothesenraum. Eine komplementäre deterministische Sichtweise charakterisiert die Generalisierung über die robustness der $\widehat{h}$: Ihre Vorhersage ändert sich um höchstens ein begrenztes Vielfaches einer Störung der Merkmale eines Datenpunkt.

Definition

Ein Wettervorhersagesystem, das auf vergangenen Daten trainiert wurde, ist nur nützlich, wenn seine Vorhersagen auf Datenpunkte genau sind, die nicht zum Trainingsdatensatz gehörten. Die Generalisierung bezieht sich auf die Fähigkeit einer gelernten Hypothese $\widehat{h}$, ähnlich genaue Vorhersagen auf Datenpunkte zu treffen, die während des Training nicht verwendet wurden. Sie ist das zentrale Ziel des machine learning (ML): das Lernen von Mustern in den Merkmale und Labels von Datenpunkte, die über das Trainingsdatensatz hinausgehen.

Die kanonische Lerntechnik ist empirische Risikominimierung (ERM): Das ML-System (ML-System) erhält $\widehat{h} \in \mathcal{H}$, indem es das empirical risk auf dem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ minimiert. Ein niedriges empirical risk auf $\mathcal{D}^{(\text{train})}$ garantiert keine genauen Vorhersagen auf Datenpunkte außerhalb von $\mathcal{D}^{(\text{train})}$. Dieselbe Frage stellt sich für Online-Lernen und Bayesian inference, wo das Trainingsdatensatz durch die bisher verarbeiteten Datenpunkte oder durch einen Prior ersetzt wird.

Die mathematische Untersuchung der Generalisierung erfordert die Formalisierung des Konzepts von Datenpunkte außerhalb des Trainingsdatensatz.

Ein weit verbreiteter Ansatz ist die Annahme eines Wahrscheinlichkeitsmodell für die Generierung von Datenpunkte, wie z.B. die Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme). Hier werden Datenpunkte als unabhängige Zufallsvariablen (ZV) mit einer identischen, festen, aber unbekannten Wahrscheinlichkeitsverteilung $\mathbb{P}$ interpretiert. Abb. 1 skizziert diesen Aufbau: Die schattierte Ellipse stellt den Bereich typischer Datenpunkte unter $\mathbb{P}$ dar, und die beiden großen gefüllten Kreise sind die Datenpunkte des Trainingsdatensatz, die ERM mit der gelernten Hypothese $\widehat{h}$ anpasst, und die Pfeile sind kleine Störungen davon. Das Risiko der $\widehat{h}$ ist der erwartete Verlust:

$$\bar{L}\big(\widehat{h}\big) = \mathbb{E}\left\{L\left(\mathbf{z}, \widehat{h}\right)\right\}.$$

Die Differenz $\bar{L}\big(\widehat{h}\big) - \widehat{L}\big(\widehat{h}|\mathcal{D}^{(\text{train})}\big)$ ist die generalization gap.

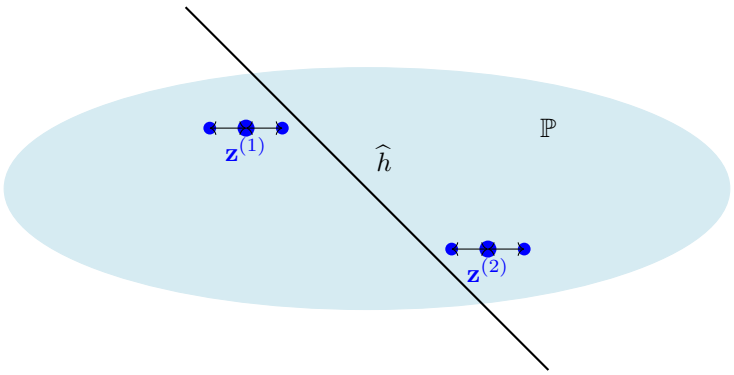


Abbildung 1: Zwei Datenpunkte $\mathbf{z}^{(1)}, \mathbf{z}^{(2)}$, die als Trainingsdatensatz verwendet werden, um eine Hypothese $\widehat{h}$ mittels ERM zu lernen. Die schattierte Ellipse markiert den Bereich typischer Datenpunkte unter einer angenommenen Wahrscheinlichkeitsverteilung $\mathbb{P}$

Für jede feste Hypothese $h \in \mathcal{H}$ ist das Risiko $\bar{L}\big(h\big)$ eine deterministische Zahl, die nur von h und $\mathbb{P}$ abhängt. Im Gegensatz dazu ist das empirical risk $\widehat{L}\big(h|\mathcal{D}^{(\text{train})}\big)$ eine ZV, die vom Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ der Größe m abhängt, welches wiederum eine Realisierung einer unabhängig und identisch verteilt (u.i.v.)-Sequenz von Zufallsvariablen $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)} \sim \mathbb{P}$ ist. Die Begrenzung der generalization gap bedeutet daher, die Wahrscheinlichkeit des (unerwünschten) Ereignis

$$\mathcal{A}^{(h)} := \left\{\mathcal{D}^{(\text{train})} : \left|\widehat{L}\big(h|\mathcal{D}^{(\text{train})}\big) - \bar{L}\big(h\big)\right| > \varepsilon\right\},$$

zu kontrollieren, wobei $\varepsilon > 0$ eine vom Benutzer festgelegte Toleranz für die generalization gap ist. Jede $h \in \mathcal{H}$ definiert ein anderes Ereignis $\mathcal{A}^{(h)}$ (siehe Abb. 2). Da die gelernte Hypothese $\widehat{h}$ von $\mathcal{D}^{(\text{train})}$ abhängt, gilt eine Schranke für die Wahrscheinlichkeit von $\mathcal{A}^{(h)}$ für eine feste h nicht für sie. Es genügt, dass $\mathcal{A}^{(h)}$ eine geringe Wahrscheinlichkeit gleichmäßig über $\mathcal{H}$ hat, das von Vapnik

and Chervonenkis (2015) eingeführte Argument: Außerhalb der Vereinigung dieser Ereignisse hat jede $h \in \mathcal{H}$, und somit auch $\widehat{h}$, eine generalization gap von höchstens ε (Shalev-Shwartz and Ben-David, 2014, Lemma 4.2; Vapnik, 2000). Abb. 2 zeigt diese Ereignisse für einen Merkmalsraum und einen Labelraum, die beide die reelle Achse sind, als Teilmengen des sample space aller Trainingsdatensätze der Größe m.

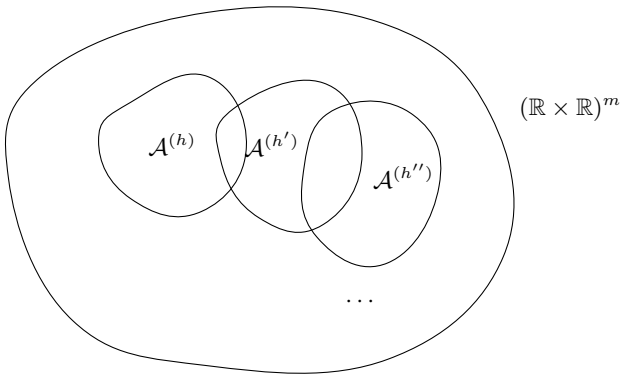


Abbildung 2: Betrachten Sie eine ML-Anwendung mit Merkmalsraum $\mathcal{X} = \mathbb{R}$ und Labelraum $\mathcal{Y} = \mathbb{R}$. Ein Trainingsdatensatz der Größe m ist ein Punkt im sample space $(\mathbb{R} \times \mathbb{R})^m$ aller möglichen Trainingsdatensätze der Größe m (der äußere Bereich). Jede Hypothese $h \in \mathcal{H}$ definiert ein Ereignis $\mathcal{A}^{(h)} \subseteq (\mathbb{R} \times \mathbb{R})^m$ (einen inneren Bereich), das jene Trainingsdatensätze sammelt, auf denen h nicht generalisiert. Damit die ML-Methode generalisiert, genügt es, dass die Vereinigung $\bigcup_{h \in \mathcal{H}} \mathcal{A}^{(h)}$ eine geringe Wahrscheinlichkeit hat; eine geringe Wahrscheinlichkeit eines einzelnen $\mathcal{A}^{(h)}$ genügt nicht, da die gelernte Hypothese vom Trainingsdatensatz abhängt.

Die gleichmäßige Kontrolle jedes Ereignis $\mathcal{A}^{(h)}$ über $\mathcal{H}$ ist nicht der einzige probabilistische Weg. Wenn die ML-Methode stabil ist, d.h. das Ersetzen eines einzelnen Datenpunkt des Trainingsdatensatz den Verlust der gelernten Hypothese im Durchschnitt nur geringfügig ändert, dann entspricht die erwartete generalization gap dieser durchschnittlichen Änderung, unabhängig von der Größe des Hypothesenraum (siehe stability; Shalev-Shwartz and Ben-David, 2014, Thm. 13.2).

Die Wahrscheinlichkeit-Theorie ist nicht der einzige Weg zur Generalisierung. Ein komplementärer deterministischer Ansatz verwendet die robustness: Eine gute Hypothese $\widehat{h}$ sollte ihre Vorhersage $\widehat{h}(\mathbf{x})$ um höchstens ein begrenztes Vielfaches der Störung des Merkmalsvektor $\mathbf{x}$ eines Datenpunkt $\mathbf{z}$ ändern, d.h. sie sollte Lipschitz-stetig sein. Zum Beispiel sollte ein object detection-System, das auf Smartphone-Fotos trainiert wurde, das Objekt immer noch erkennen, wenn ein einzelnes Pixel geändert wird. Es wurde gezeigt, dass trainierte tiefe Netze diese Eigenschaft nicht besitzen (Su et al., 2019). Das System sollte auch das gleiche Ergebnis liefern, wenn das Objekt im Bild leicht verschoben

oder verformt wird (Mallat, 2016, Sect. 3). Die Pfeile in Abb. 1 zeigen kleine Störungen jedes Datenpunkt im Trainingsdatensatz einer ML-Methode an.
Siehe auch: generalization gap, Risiko, empirical risk, u.i.v.-Annahme, Trainingsdatensatz, Überanpassung, Regularisierung, stability.

Querverweise

Literatur

1. Mallat (2016). Understanding deep convolutional networks. Philos. Trans. Roy. Soc. A, 374(2065), pp. Art. no. 20150203.

2. Su et al. (2019). One Pixel Attack for Fooling Deep Neural Networks. IEEE Trans. Evol. Comput., 23(5), pp. 828–841.

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

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

5. Vapnik and Chervonenkis (2015). On the uniform convergence of relative frequencies of events to their probabilities. In: Measures of Complexity: Festschrift for Alexey Chervonenkis, pp. 11–30. Springer.