Regularisierung

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

Regularisierung bezieht sich auf Änderungen an einer machine learning (ML)-Methode, die ihre Generalisierung verbessern. Drei Wege werden danach unterschieden, welche Komponente der Methode modifiziert wird: das Beschneiden des Hypothesenraum, das Hinzufügen eines Strafterm zur Verlustfunktion und das Erweitern des Trainingsdatensatz mit gestörten Kopien seiner Datenpunkte. Die drei Wege führen oft zur gleichen gelernten Hypothese. Zum Beispiel ist das Erweitern des Trainingsdatensatz der lineare Regression mit Gaußschen Merkmal-Störungen äquivalent zum Hinzufügen des Ridge-Regression Strafterm zum Trainingsfehler.

Definition

Training eines großen tiefes Netz (wie eines großes Sprachmodell (LLM)) auf einem festen Trainingsdatensatz mittels einfacher empirische Risikominimierung (ERM) führt typischerweise zu Überanpassung: Die gelernte Hypothese $\widehat{h}$ funktioniert gut auf dem Trainingsdatensatz, aber schlecht außerhalb davon. Regularisierung bezieht sich auf Änderungen an einer machine learning (ML)-Methode, um sicherzustellen, dass die gelernte Hypothese auf Datenpunkte außerhalb des Trainingsdatensatz nahezu genauso gut funktioniert wie auf denen innerhalb (Goodfellow et al., 2016, Kap. 7).
Für ERM-basierte Methoden kann die Regularisierung auf drei Arten erfolgen:

1) Model pruning: Verkleinern des Hypothesenraum $\mathcal{H}$ (der Menge der Kandidaten-Hypothesen h) zu einem kleineren $\mathcal{H}'$. Für ein parametrisches Modell kann die Verkleinerung durch Beschränkungen der Modellparameter implementiert werden (z. B. $w_1 \in [0.4, 0.6]$ für das Gewicht des Merkmal x_1 in der lineare Regression).

2) Verlust-Penalisierung: Hinzufügen eines Strafterm zum Trainingsfehler der ERM. Für die squared error loss ist der Strafterm eine Schätzung, um wie viel das Risiko den durchschnittlichen Verlust auf dem Trainingsdatensatz übersteigt (Hastie et al., 2009, Abschn. 7.4); für andere Verluste steuert er die generalization gap, ohne sie zu schätzen (Shalev-Shwartz and Ben-David, 2014, Kor. 13.6) (siehe Stützvektormaschine (SVM)).

3) Datenerweiterung: Erweitern des Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ mit gestörten Kopien seiner Datenpunkte (z. B. Hinzufügen der Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) zum Merkmalsvektor jedes Datenpunkt).

Abb. 1 veranschaulicht die drei Wege.

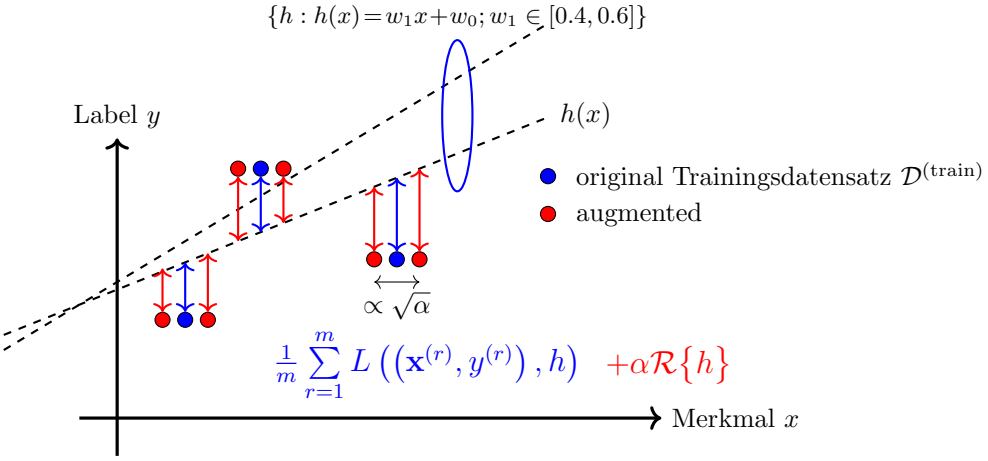

Abbildung 1: Drei Wege zur Regularisierung einer ERM-basierten Methode, dargestellt für die lineare Regression. 1) Model pruning: Das Beschränken der Modellparameter (hier $w_1 \in [0.4, 0.6]$) verkleinert den Hypothesenraum auf die Hypothesen innerhalb der blauen Ellipse. 2) Verlust-Penalisierung: Der rote Strafterm $\alpha \mathcal{R}\{h\}$ wird zum blauen Trainingsfehler nahe dem unteren Rand des Diagramms hinzugefügt. 3) Datenerweiterung: Die ursprünglichen Datenpunkte (blau) werden mit gestörten Kopien (rot) erweitert, die um einen Betrag $\propto \sqrt{\alpha}$ verschoben sind (Doppelpfeil für die rechteste Gruppe gezeigt).

Diese Wege können zur gleichen gelernten Hypothese führen. Betrachten wir als Beispiel die Datenerweiterung für die lineare Regression, die Störungen mit Null-Mittelwert und Kovarianzmatrix $\sigma^2 \mathbf{I}$ zum Merkmalsvektor jedes Datenpunkt hinzufügt. Asymptotisch in der Anzahl der Störungen stimmt die resultierende gelernte Hypothese mit derjenigen überein, die aus der Ridge-Regression erhalten wird (Bishop, 1995; Goodfellow et al., 2016, Abschn. 7.5). Die Ridge-Regression fügt den Strafterm $\sigma^2 \|\mathbf{w}\|_2^2$ zum Trainingsfehler der lineare Regression hinzu (Hastie et al., 2009, Abschn. 3.4.1). Gemäß der Lagrangian-Dualität ist das Hinzufügen dieses Strafterm äquivalent zur Verkleinerung des Hypothesenraum auf die Menge der Hypothesen mit $\|\mathbf{w}\|_2^2 \leq C$ für eine Konstante C (Hastie et al., 2009, Gl. (3.42)).

Die oben genannten Regularisierungstechniken sind nicht auf ERM-basierte Methoden beschränkt, sondern können auch auf andere Arten von ML-Algorithmen angewendet werden. Zum Beispiel kann beim Online-Lernen

die Follow-The-Leader (FTL)-Methode durch Hinzufügen eines Regularisierungsterm zur rundenbasierten Zielfunktion der FTL regularisiert werden. Die resultierende Follow-The-Regularized-Leader (FTRL)-Methode erreicht einen geringeren Regret als die einfache FTL (Hazan, 2022, Kap. 5). Ein weiteres Beispiel für Regularisierung ist die Verwendung einer prior distribution über $\mathcal{H}$ in der Bayesian inference. Die prior distribution wirkt als Regularisierungsterm: Der negative Logarithmus der prior distribution ist ein Strafterm, so dass die Maximum-a-posteriori-Schätzung der Modellparameter der lineare Regression unter einer Gaußschen prior distribution die Ridge-Regression-Lösung ist (Bishop, 2006, Abschn. 1.2.5).

Siehe auch: Überanpassung, Generalisierung, model pruning, Strafterm, Datenerweiterung, Ridge-Regression, least absolute shrinkage and selection operator (Lasso).

Querverweise

Literatur

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

2. Bishop (1995). Training with Noise is Equivalent to Tikhonov Regularization. Neural Computation, 7(1), pp. 108–116.

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

4. Hazan (2022). Introduction to Online Convex Optimization. 2nd ed., MIT Press, Cambridge, MA, USA.

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

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