Datenerweiterung

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

Datenerweiterung erweitert einen Trainingsdatensatz mit synthetischen Datenpunkte, die durch Transformationen oder Störungen bestehender Datenpunkte gewonnen werden. Typische Beispiele sind das Drehen oder Beschneiden eines Bildes und das Hinzufügen kleiner Störungen zu den Merkmale und Labels eines gemessenen Datenpunkts. Wenn sich das Label unter einer Transformation nicht ändert, führt das Training mit transformierten Kopien dazu, dass die gelernte Hypothese unter dieser Transformation wenig variiert. Sensorgeräte messen mit begrenzter Präzision, sodass gestörte Kopien das uncertaintysintervall um jeden aufgezeichneten Wert widerspiegeln. Datenerweiterung ist eine Form der Regularisierung, neben model pruning und dem Hinzufügen eines Strafterms zum Trainingsfehler. Für die lineare Regression ist das Stören der Merkmale mit Rauschen mit Mittelwert Null, dessen Kovarianzmatrix ein Vielfaches der Einheitsmatrix ist, im Erwartungswert über das Rauschen äquivalent zum Hinzufügen des Ridge-Regression-Strafterms.

Definition

Drei gemessene Tage in Krems an der Donau (Österreich) sind alles, woraus ein Prognostiker lernen kann. Jeder Tag ist ein Datenpunkt, gekennzeichnet durch die Lufttemperatur um 06:00 (UTC) als sein Merkmal x und die Temperatur um 15:00 (UTC) desselben Tages als sein Label y (Abb. 1). Die Abbildung zeigt auch eine gelernte Hypothese $\widehat{h}$, die durch empirische Risikominimierung (ERM) unter Verwendung von Polynomen zweiten Grades als Hypothesenraum erhalten wurde.

Die gelernte Hypothese erreicht einen Trainingsfehler von Null, liefert aber schlechte Vorhersagen außerhalb des Trainingsdatensatzs, was Überanpassung ist. Auf einem zurückgehaltenen Satz von 106 Datenpunkte, die von anderen Tagen des Jahres 2024 stammen und deren Merkmal innerhalb des gemessenen Bereichs liegt, beträgt ihr durchschnittlicher squared error loss 335.3 Quadratgrad. Ein Grund für das Versagen von $\widehat{h}$ ist, dass der Trainingsdatensatz zu klein ist. Für fast jeden Trainingsdatensatz der Stichprobengröße $m = 3$ gibt es ein Polynom zweiten Grades, das durch seine Datenpunkte verläuft.

Datenerweiterung erweitert einen Datensatz mit synthetischen Datenpunkte, die durch Störungen der Merkmale und Labels der gemessenen Datenpunkte ge-

wonnen werden. Sensorgeräte weisen Unvollkommenheiten wie thermisches Rauschen auf, sodass ein Messwert x ein Repräsentant eines uncertaintysintervalls $\mathcal{U}^{(x)}$ ist. Jeder andere Wert aus diesem Intervall hätte stattdessen beobachtet werden können.

Betrachtet man die Wetterdaten aus Abb. 1, so trägt ein Temperaturmesswert x ein $\mathcal{U}^{(x)} = [x - 1, x + 1]$. Jeder gemessene Tag (x, y) wird durch 400 Kopien $(x + \varepsilon, y + \varepsilon')$ ersetzt, deren Störungen ε und ε' Realisierungen von Zufallsvariablen (ZV) sind, die gleichmäßig auf $[-1, 1]$ verteilt sind. Der erweiterte Trainingsdatensatz $\widetilde{\mathcal{D}}$ enthält 1200 Datenpunkte, von denen keiner eine neue Temperaturmessung erforderte. Die Breite des uncertaintysintervalls, das für die Störung der Merkmale und Labels verwendet wird, ist ein Hyperparameter. Eine grundlegende Methode, dies zu bestimmen, besteht darin, einige Optionen auszuprobieren und den Validierungsfehler der resultierenden gelernten Hypothesen zu vergleichen.

Abb. 1 zeigt auch das Polynom zweiten Grades $\widehat{h}'$, das ERM unter Verwendung von $\widetilde{\mathcal{D}}$ als Trainingsdatensatz lernt. Die Hypothese $\widehat{h}'$ verläuft nicht mehr durch die drei gemessenen Tage, liefert aber bessere Vorhersagen außerhalb des Trainingsdatensatzs. Ihr Trainingsfehler auf diesen beträgt 4.0 Quadratgrad, aber ihr durchschnittlicher squared error loss auf den zurückgehaltenen Datenpunkte sinkt auf 25.7. Der Koeffizient von x^2 fällt von $\widehat{w}_2 = 1.87$ für $\widehat{h}$ auf $\widehat{w}'_2 = 0.11$ für $\widehat{h}'$.

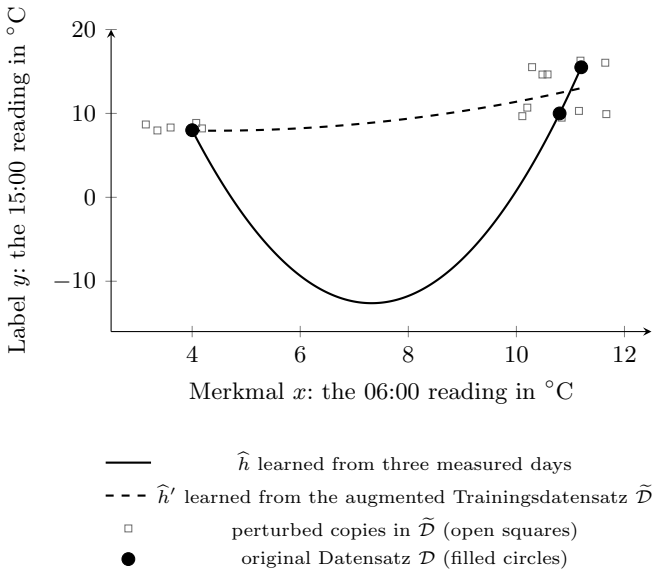

Abbildung 1: Drei Datenpunkte (gefüllte Kreise), die die täglichen Wetterbedingungen in Krems an der Donau darstellen. Die durchgezogene Kurve ist das Polynom zweiten Grades $\widehat{h}$ mit einem Trainingsfehler von Null auf diesen. Sie fällt weit unter jede gemessene Temperatur. Die offenen Quadrate sind Kopien, die aus dem uncertaintysintervall jedes gemessenen Tages gezogen wurden, fünf Kopien von jedem Tag. Die gestrichelte Kurve ist die Hypothese $\widehat{h}'$, die aus ihnen gelernt wurde. Daten generiert von `pythondemos/dataaug.py`

Das Erweitern von $\mathcal{D}$ mit gestörten Kopien ist eine grundlegende Form der Regularisierung, neben model pruning, das den Hypothesenraum verkleinert, und dem Hinzufügen eines Strafterms zum Trainingsfehler. Diese verschiedenen Formen der Regularisierung sind manchmal in ihrer Wirkung auf die gelernte Hypothese äquivalent (Abb. 2). Für die lineare Regression führt das Hinzufügen von Störungen mit Mittelwert Null und Kovarianzmatrix $\sigma^2\mathbf{I}$ zu jedem Merkmalsvektor zur Ridge-Regression-Lösung, asymptotisch in der Anzahl der Kopien. Ihr Strafterm ist $\sigma^2\|\mathbf{w}\|_2^2$ für den Gewicht-Vektor $\mathbf{w}$ (Bishop, 1995; Goodfellow et al., 2016, Abschn. 7.5).

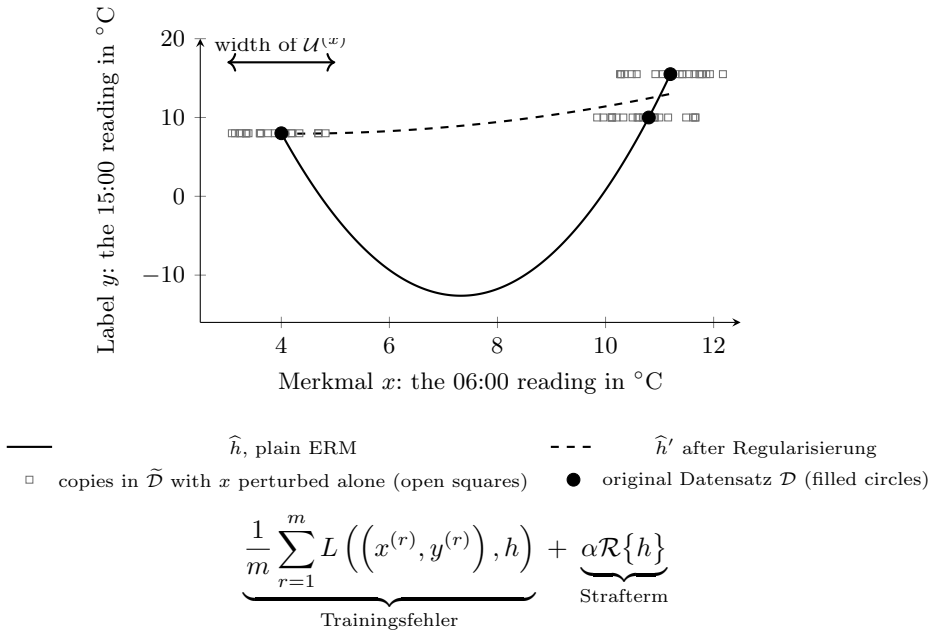

Abbildung 2: Die drei Formen der Regularisierung auf den Datenpunkte von Abb. 1. Hier wird nur das Merkmal gestört, der Fall, den die Äquivalenz mit der Ridge-Regression abdeckt, sodass jedes offene Quadrat neben seinem gemessenen Tag mit demselben Label liegt. Der Doppelpfeil markiert die Breite des uncertaintysintervalls, aus dem sie gezogen werden. Das Erweitern des Trainingsdatensatzs mit ihnen ersetzt die durchgezogene Kurve $\widehat{h}$ durch die gestrichelte, die innerhalb von 0.05 °C der $\widehat{h}'$ aus Abb. 1 bleibt. Das Hinzufügen des Strafterms $\alpha\mathcal{R}\{h\}$ zum Trainingsfehler, unterhalb des Diagramms gezeigt, ist die zweite Form. Das Verkleinern des Hypothesenraums auf $\{h : |w_2| \leq 0.10\}$ ist die dritte; die Hypothese, die sie liefert, liegt innerhalb von 0.11 °C der gestrichelten Kurve und wird nicht separat gezeichnet. Daten generiert von `pythondemos/dataaug.py`

Anstatt Rauschen zu den Merkmale und Labels hinzuzufügen, können Störungen auch von einer Transformation $\mathcal{T}$ des Merkmalsraums herrühren. Eine Transformation, von der bekannt ist, dass sie das Label unverändert lässt, ist eine Symmetrie der Lernaufgabe. Ein gedrehtes oder beschnittenes Bild einer Katze zeigt immer noch eine Katze, während ihr Merkmalsvektor, das durch stacking der Pixelfarbintensitäten erhalten wird, sich ändert. Die Transformation muss das Label erhalten, und eine Rotation tut dies nur bis zu einem gewissen Grad. Abb. 3 reduziert Bilder der handgeschriebenen Ziffern 6 und 9 auf ihre ersten beiden Hauptrichtungen (siehe Hauptkomponentenanalyse (PCA)) als Merkmale x_1, x_2. Eine kleine Rotation $\mathcal{T}$ hält ein Bild einer 6 unter den anderen 6ern. Das Drehen eines Bildes um 180 Grad führt es jedoch zu den 9ern, wobei sich das Label geändert hat.

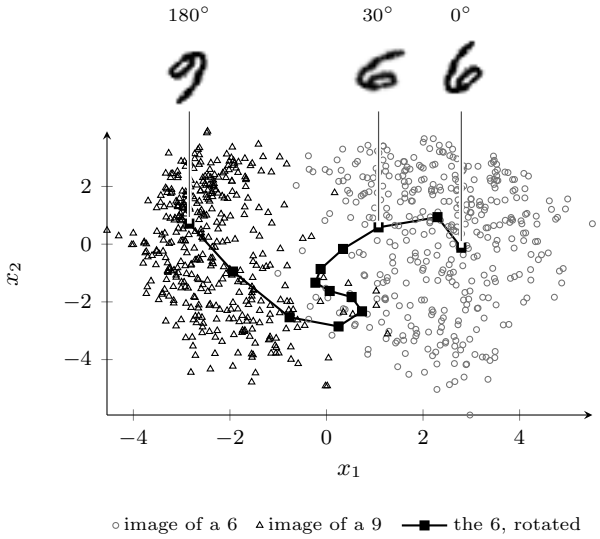

Abbildung 3: Bilder der handgeschriebenen Ziffern 6 (Kreise) und 9 (Dreiecke), jeweils dargestellt durch zwei Merkmale x_1, x_2, die aus PCA gewonnen wurden. Die durchgezogene Linie mit quadratischen Markierungen ist der Pfad, den ein Bild einer 6 verfolgt, wenn es von 0 bis 180 Grad gedreht wird. Die drei Einlagen über dem Diagramm zeigen dieses Bild aufrecht, um 30 Grad gedreht und um 180 Grad gedreht, jeweils über dem entsprechenden Punkt des Pfades platziert. Ihre ersten Schritte halten das Bild unter den 6ern, sodass eine Rotation um einen kleinen Winkel das Label erhält. Ihre letzten Schritte platzieren das Bild unter den 9ern. Daten generiert von `pythondemos/dataaug.py`

Siehe auch: Regularisierung, model pruning, Strafterm, Ridge-Regression, Überanpassung, ERM, Trainingsdatensatz, Datensatz, Hypothesenraum, Hypothese, Datenpunkt, Merkmal, Label, Merkmalsvektor, Merkmalsraum, uncertainty, Hyperparameter, Validierungsfehler, polynomial regression, Lernaufgabe.

Querverweise

Literatur

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

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