hochdimensionales Regime

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

Eine maschinelles Lernen (ML)-Methode arbeitet im hochdimensionalen Regime, wenn die effektive Dimension ihres Hypothesenraums den Stichprobengröße m des Trainingsdatensatzs übersteigt. Für ein lineares Modell beginnt das Regime dort, wo die Anzahl d der Merkmale die Anzahl m der Datenpunkte überschreitet. Die Merkmalsmatrix hat dann höchstens den Rang m, sodass die Lösung der kleinste Quadrate nicht eindeutig ist und jede Lösung den Trainingsdatensatz interpoliert. Ein Trainingsfehler von null trägt daher keine Information über das Risiko bei, sodass das Regime den üblichen Anhaltspunkt für die Generalisierung beseitigt. Regularisierung und Dimensionsreduktion stellen eine brauchbare Methode wieder her, erstere durch Hinzufügen eines Strafterms zum Trainingsfehler und letztere durch Verkleinerung des Hypothesenraums. Die hochdimensionale Statistik entwickelt Theorie und Methoden für dieses Regime, und ihre Garantien erfordern eine strukturelle Annahme wie etwa Dünnbesetztheit.

Definition

Betrachtet sei ein Wetterdienst in Krems an der Donau (Österreich), der die Lufttemperatur eines Tages um 15:00 Uhr (UTC) aus den Messwerten um 15:00 Uhr der vorangegangenen d Tage vorhersagt. Jeder Tag ist ein Datenpunkt mit dem Merkmalsvektor $\mathbf{x} \in \mathbb{R}^d$ und dem Label $y \in \mathbb{R}$, und ein lineares Modell $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ wird mittels der kleinste Quadrate auf einem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ aus $m = 30$ Tagen des Jahres 2023 gelernt. Ein Tag wird nur dann in eine der beiden Mengen aufgenommen, wenn alle 60 seiner Vorgänger aufgezeichnet sind, was durch die zwei Tage, die in der stündlichen Aufzeichnung fehlen, den 31. Dezember 2023 und den 12. Januar 2024, für die ersten Wochen des Jahres 2024 ausgeschlossen ist; das Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ besteht daher aus den 294 Tagen des Jahres 2024 ab dem 13. März. Für $d = 1$ beträgt der Trainingsfehler 17.9 und der Validierungsfehler 14.7 Grad zum Quadrat. Für $d = 29$ lauten die beiden Werte 0.05 und 501, und für $d = 30$ betragen sie 0 und 955. Sobald die Anzahl der Merkmale die Anzahl der Datenpunkte in $\mathcal{D}^{(\text{train})}$ erreicht, liefert der Trainingsfehler keine Information über den Verlust an Tagen außerhalb von $\mathcal{D}^{(\text{train})}$. Abb. 1 zeigt beide Fehler für jedes d von 1 bis 60.

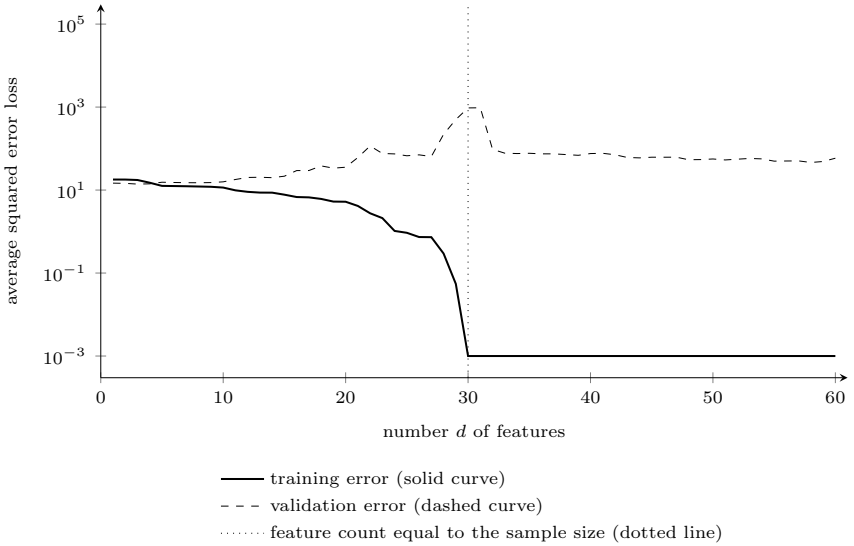


Abbildung 1: Der Trainingsfehler (durchgezogene Kurve) und der Validierungsfehler (gestrichelte Kurve) eines lineares Modells, das die Lufttemperatur eines Tages in Krems um 15:00 Uhr (UTC) aus den Messwerten um 15:00 Uhr der vorangegangenen d Tage vorhersagt, gelernt mittels der kleinste Quadrate auf $m = 30$ Tagen des Jahres 2023 und ausgewertet auf den 294 Tagen des Jahres 2024, die über 60 aufeinanderfolgende Vorgänger verfügen. Der Trainingsfehler ist überall dort, wo er null beträgt, am unteren Achsenrand 10^{-3} eingezeichnet. Die gepunktete Linie markiert die Merkmalsanzahl $d = m = 30$, die letzte vor dem hochdimensionalen Regime, welches dort beginnt, wo d den Wert m überschreitet. Daten erzeugt mit `pythondemos/highdimregime.py`

Eine maschinelles Lernen (ML)-Methode arbeitet im hochdimensionalen Regime, wenn die effektive Dimension ihres Hypothesenraums $\mathcal{H}$ den Stichprobengröße m des Trainingsdatensatzs übersteigt,

$$d_{\text{eff}}\left(\mathcal{H}\right) > m. \tag{1}$$

Für das obige lineares Modell lautet die Bedingung (1) $d > m$. Die empirische Risikominimierung (ERM) mit dem quadratischer Verlust liefert eine Lösung des Problems der kleinste Quadrate

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{m}\sum_{r=1}^{m}\left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right)^2. \tag{2}$$

Hierbei fasst die Merkmalsmatrix $\mathbf{X} \in \mathbb{R}^{m\times d}$ die Merkmalsvektoren aus $\mathcal{D}^{(\text{train})}$ als ihre Zeilen zusammen und der Vektor $\mathbf{y} \in \mathbb{R}^m$ deren Labels.

Der Rang von $\mathbf{X}$ beträgt höchstens m, sodass für $d > m$ der Nullraum von $\mathbf{X}$ einen vom Nullvektor verschiedenen Vektor enthält und die Lösung des Problems der kleinste Quadrate (2) nicht eindeutig ist: Seine Minimierer bilden eine

affine Menge der Dimension von mindestens $d-m$. Die Pseudoinverse wählt denjenigen mit der kleinsten euklidische Norm aus, $\widehat{\mathbf{w}} = \mathbf{X}^{+}\mathbf{y}$. Jeder Minimierer drückt den Trainingsfehler auf null, was einer Interpolation von $\mathcal{D}^{(\text{train})}$ entspricht. Der Fluch der Dimensionalität ist eine gesonderte Folge eines großen d: Die Datenpunkte verteilen sich dünn über den Merkmalsraum $\mathbb{R}^d$, unabhängig davon, wie viele es sind, während das hochdimensionale Regime zwei Anzahlen miteinander vergleicht.

Generalisierung ist die Güte einer gelernten Hypothese auf Datenpunkte außerhalb von $\mathcal{D}^{(\text{train})}$, und das Regime beseitigt den Anhaltspunkt, den ein kleiner Trainingsfehler normalerweise dafür liefert. Eine Hypothese, die $\mathcal{D}^{(\text{train})}$ interpoliert, weist einen Trainingsfehler von null auf, unabhängig davon, welche Labels die Datenpunkte von $\mathcal{D}^{(\text{train})}$ tragen, sodass ein Trainingsfehler von null das Risiko unbestimmt lässt und Überanpassung nicht allein anhand von $\mathcal{D}^{(\text{train})}$ diagnostiziert werden kann. Ein großes künstliches neuronales Netz (KNN) wird im selben Regime trainiert, mit weitaus mehr trainierbaren Modellparameter als Datenpunkte (Zhang et al., 2017, Sect. 1). Ein tiefes Netz erreicht selbst auf einem Trainingsdatensatz, dessen Labels durch zufällige ersetzt wurden, einen Trainingsfehler von null (Zhang et al., 2017, Sect. 2). Die Lerntheorie misst die Größe von $\mathcal{H}$ durch seine Vapnik-Chervonenkis-Dimension (VC-Dimension) und durch seine Rademacher-Komplexität. Der Stichprobengröße, den eine Garantie für die Generalisierung benötigt, wächst proportional zur VC-Dimension (Shalev-Shwartz and Ben-David, 2014, Thm. 6.8).

Zwei Wege stellen eine brauchbare Methode auf denselben Merkmale wieder her, und beide sind Formen der Regularisierung. Das Hinzufügen des Strafterms $\alpha\|\mathbf{w}\|_2^2$ zum Trainingsfehler ergibt die Ridge-Regression. Mit $\alpha = 100$ bleibt der Validierungsfehler des Wetterbeispiels für jedes d von 1 bis 60 zwischen 13.8 und 32.7 Grad zum Quadrat. Dimensionsreduktion verkleinert stattdessen $\mathcal{H}$: Die kleinste Quadrate auf den ersten $d' = 5$ Hauptrichtungen der Merkmalsvektoren von $\mathcal{D}^{(\text{train})}$, was der Hauptkomponentenanalyse (PCA) entspricht, hält den Validierungsfehler zwischen 13.9 und 23.2. Bei $d = m$ interpoliert keine der beiden gelernten Hypothesen den Trainingsdatensatz $\mathcal{D}^{(\text{train})}$: Der Trainingsfehler beträgt 5.1 mit dem Strafterm und 15.9 nach der Reduktion. Abb. 2 stellt den Validierungsfehler beider Verfahren dem Validierungsfehler der kleinste Quadrate auf allen d Merkmale gegenüber.

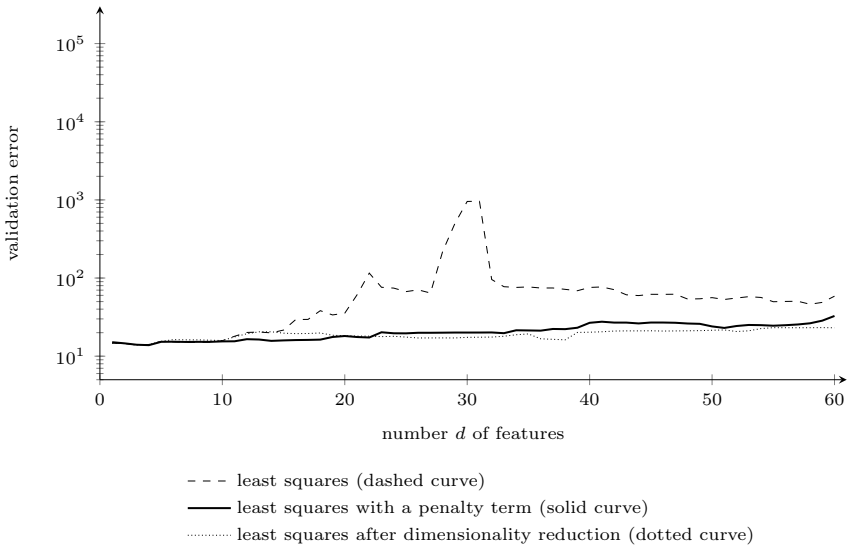


Abbildung 2: Der Validierungsfehler desselben lineares Modells in Krems in Abhängigkeit von der Anzahl d der Merkmale, für die kleinste Quadrate (gestrichelte Kurve), für die kleinste Quadrate mit dem Strafterm bei $\alpha = 100$ (durchgezogene Kurve) und für die kleinste Quadrate auf den ersten $d' = 5$ Hauptrichtungen (gepunktete Kurve). Daten erzeugt mit `pythondemos/highdimregime.py`

Die hochdimensionale Statistik ist das Teilgebiet der mathematischen Statistik, das Theorie und Methoden für dieses Regime entwickelt (Bühlmann and Geer, 2011; Wainwright, 2019, Sect. 1.1). Ohne eine strukturelle Annahme über die Modellparameter schätzt keine Methode diese konsistent, sobald d den Wert m übersteigt (Wainwright, 2019, Sect. 7.1.1). Eine solche Annahme ist die Dünnbesetztheit, die besagt, dass nur wenige Modellparameter von null verschieden sind. Das Lasso (Lasso) nutzt dies mit einem aus der ℓ_1-Norm gebildeten Strafterm aus (Tibshirani, 1996). Das Regime ist überall dort verbreitet, wo Merkmale günstig zu gewinnen sind und Datenpunkte nicht: Ein Genexpressions-Datensatz mit 2308 Merkmale und 63 Datenpunkte erfordert Regularisierung, bevor irgendein Klassifikator darauf gelernt werden kann (Hastie et al., 2009, Sect. 18.2).

Siehe auch: effektive Dimension, Fluch der Dimensionalität, Interpolation, Überanpassung, Generalisierung, Regularisierung, Dimensionsreduktion, Ridge-Regression, Lasso, VC-Dimension.

Querverweise

Literatur

1. Bühlmann and Geer (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer-Verlag, Berlin, Germany.

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

3. Tibshirani (1996). Regression Shrinkage and Selection via the Lasso. J. Roy. Statist. Soc.: Ser. B (Methodological), 58(1), pp. 267–288.

4. Wainwright (2019). High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge Univ. Press, Cambridge, U.K.

5. Zhang et al. (2017). Understanding deep learning requires rethinking generalization. In: Int. Conf. Learn. Represent. (ICLR).

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.