Ridge-Regression

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 Ridge-Regression lernt eine lineare Hypothese-Abbildung, indem sie den durchschnittlichen squared error loss auf einem Trainingsdatensatz plus einen Strafterm minimiert, der proportional zum Quadrat der Euclidean norm der Modellparameter ist. Der Strafterm erzwingt Regularisierung: Er verhindert Überanpassung, wenn die Anzahl der Merkmale die Anzahl der Datenpunkte übersteigt oder wenn Merkmale nahezu kollinear sind. Im Gegensatz zu least absolute shrinkage and selection operator (Lasso) verwendet die Ridge-Regression eine Zielfunktion mit einem eindeutigen Minimierer, der für jedes Trainingsdatensatz in geschlossener Form durch ein System von lineare Gleichungen gegeben ist. In der Basis der singulären Vektoren der Merkmalsmatrix schrumpft die Ridge-Regression jede Komponente der lineare Regression-Lösung um einen Faktor zwischen 0 und 1, der für einen kleinen Singulärwert klein ist, was die Rauschverstärkung der lineare Regression begrenzt. Der Strafterm entspricht der Erhöhung des durchschnittlichen squared error loss unter Datenerweiterung mit zufälligen Merkmal-Störungen. Ridge-Regression auf transformierten Merkmalsvektoren, die aus einem Kernel konstruiert werden, ist Kernel-Ridge-Regression.

Definition

Die Vorhersage des 15:00-Werts (UTC) von morgen in Krems an der Donau anhand der 24 Stundentemperaturen von heute ist eine machine learning (ML)-Anwendung. Ein Datenpunkt ist ein Tag, sein Merkmalsvektor $\mathbf{x} \in \mathbb{R}^{24}$ listet die Stundentemperaturen des Tages auf, und sein Label y ist der 15:00-Wert (UTC) des nächsten Tages (GeoSphere Austria, 2026). Ein Trainingsdatensatz von $m = 20$ Tagen im Juli 2024 enthält weniger Datenpunkte als Merkmale, $d = 24$, und die Merkmale sind nahezu kollinear: Die Temperaturen um 13:00 und um 14:00 haben eine Korrelation von 0.93. Lineare Regression lernt dann ein lineares Modell, das die 20 Training-Labels exakt reproduziert und die 15:00-Werte des August 2024, das Testdatensatz, mit einem mean squared error (MSE) von 104 vorhersagt. Die Ridge-Regression lernt aus demselben Trainingsdatensatz ein lineares Modell mit einem Training-MSE von 7.5 und einem MSE von 20.1 auf dem Testdatensatz (siehe Abb. 1).

Die Ridge-Regression lernt die Modellparameter $\mathbf{w}$ einer linearen Hypothese-Abbildung $h^{(\mathbf{w})}(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$, genauso wie lineare Regression. Sie minimiert den penalisierten durchschnittlichen squared error loss auf einem Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}$ (Hoerl and Kennard, 1970),

$$\widehat{\mathbf{w}}^{(\alpha)} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right)^2 + \alpha \|\mathbf{w}\|_2^2. \tag{1}$$

Das Optimierungsproblem lässt sich kompakter mit der Merkmalsmatrix $\mathbf{X}$ und dem label vector $\mathbf{y}$ schreiben,

$$\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)^\top \in \mathbb{R}^{m \times d}, \qquad \mathbf{y} = \left(y^{(1)}, \ldots, y^{(m)}\right)^\top \in \mathbb{R}^m.$$

In Bezug auf $\mathbf{X}$ und $\mathbf{y}$ lautet das Ridge-Regression-Problem (1)

$$\widehat{\mathbf{w}}^{(\alpha)} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{m} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \alpha \|\mathbf{w}\|_2^2.$$

Die Ridge-Regression entsteht aus der lineare Regression, indem ein Strafterm zum durchschnittlichen squared error loss hinzugefügt wird. Der Strafterm ist die skalierte quadratische Euclidean norm $\alpha \|\mathbf{w}\|_2^2$ mit einem Regularisierung-Parameter $\alpha > 0$ (Hastie et al., 2009, Sekt. 3.4.1, Gl. (3.41)). Das Ergebnis ist eine Instanz von regularized empirical risk minimization (RERM). Der Zweck des Strafterm ist Regularisierung. Er verhindert Überanpassung im high-dimensional regime, wo die Anzahl der Merkmale d die Anzahl der Datenpunkte m im Trainingsdatensatz übersteigt (Hastie et al., 2009, Sekt. 18.1), wie es bei den 20 Tagen mit jeweils 24 Stundentemperaturen der Fall ist. Eine alternative Methode zur Regularisierung ist das least absolute shrinkage and selection operator (Lasso), das sich von der Ridge-Regression in der Konstruktion des Strafterm unterscheidet: Lasso verwendet die ℓ_1-Norm $\|\mathbf{w}\|_1$ statt der quadratischen Euclidean norm $\|\mathbf{w}\|_2^2$.

Im Gegensatz zu Lasso, dessen Zielfunktion mehrere Minimierer haben kann, besitzt die Ridge-Regression-Zielfunktion für jede Merkmalsmatrix $\mathbf{X}$ und label vector $\mathbf{y}$ den eindeutigen Minimierer

$$\widehat{\mathbf{w}}^{(\alpha)} = \left(\mathbf{X}^\top \mathbf{X} + \alpha m \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{2}$$

Die Normalgleichungen der gewöhnlichen lineare Regression können im Vergleich dazu unendlich viele Lösungen haben, wie bei den 20 Tagen im Juli: Jede davon reproduziert die Trainings-Labels. Training ist eine einmalige Berechnung: Die Ridge-Regression benötigt keine iterative Optimierungsverfahren und damit keine Fixpunktinterpretation, sondern nur die Lösung des $d \times d$-Systems von lineare Gleichungen $\left(\mathbf{X}^\top \mathbf{X} + \alpha m \mathbf{I}\right)\mathbf{w} = \mathbf{X}^\top \mathbf{y}$ mit einer invertible Matrix, der zero-gradient condition der penalisierten Zielfunktion. Bei vielen Merkmale ist Gradientenabstieg auf der penalisierten Zielfunktion die Alternative, und der Strafterm addiert αm zu jedem Eigenwert von $\mathbf{X}^\top \mathbf{X}$, was die condition number senkt, die die Konvergenzrate von Gradientenabstieg bestimmt.

Die Singulärwertzerlegung (SVD) der Merkmalsmatrix zeigt, warum die lineare Regression- Lösung das Rauschen in den Labels verstärkt und wie der Strafterm die Verstärkung begrenzt. Schreibe $\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^\top$ mit Singulärwerte $\lambda_1 \geq \ldots \geq \lambda_p > 0$, wobei p der rank von $\mathbf{X}$ ist, mit linken singulären Vektoren $\mathbf{v}^{(j)}$ und rechten singulären Vektoren $\mathbf{u}^{(j)}$. Die Lösung der lineare Regression mit der kleinsten Euclidean norm ist

$$\widehat{\mathbf{w}} = \sum_{j=1}^{p} \frac{(\mathbf{v}^{(j)})^\top \mathbf{y}}{\lambda_j} \mathbf{u}^{(j)}.$$

Zwei Eigenschaften dieser Lösung verursachen ihren großen Testfehler. Eine Komponente von $\mathbf{y}$ entlang $\mathbf{v}^{(j)}$ wird durch λ_j geteilt, sodass ein kleiner Singulärwert das Rauschen in $\mathbf{y}$ um den Faktor $1/\lambda_j$ verstärkt. Kleine Singulärwerte entstehen durch nahezu kollineare Merkmale: Die Singulärwerte der 20 Tage mit Stundentemperaturen reichen von 504 bis 0.21. Und eine Richtung $\mathbf{u}^{(j)}$ mit $j > p$, die immer existiert, wenn $d > m$, lässt $\mathbf{X}\mathbf{w}$ unverändert, sodass das Trainingsdatensatz die Komponente von $\mathbf{w}$ entlang dieser Richtung nicht bestimmt. Setzt man die SVD in die geschlossene Form (2) ein, ergibt sich (Hastie et al., 2009, Sekt. 3.4.1, Gl. (3.47))

$$\widehat{\mathbf{w}}^{(\alpha)} = \sum_{j=1}^{p} \frac{\lambda_j^2}{\lambda_j^2 + \alpha m} \cdot \frac{(\mathbf{v}^{(j)})^\top \mathbf{y}}{\lambda_j} \mathbf{u}^{(j)}.$$

Die Ridge-Regression multipliziert jede Komponente von $\widehat{\mathbf{w}}$ in der Basis der rechten singulären Vektoren mit dem Shrinkage-Faktor $\lambda_j^2/(\lambda_j^2 + \alpha m)$, der strikt zwischen 0 und 1 liegt. Für λ_j^2, die viel größer als αm sind, ist der Faktor nahe bei 1 und die Komponente wird beibehalten. Für λ_j^2, die viel kleiner als αm sind, wird die Komponente gegen 0 geschrumpft. Das Produkt aus dem Faktor und $1/\lambda_j$ ist $\lambda_j/(\lambda_j^2 + \alpha m)$, das niemals $1/(2\sqrt{\alpha m})$ übersteigt, sodass die Rauschverstärkung begrenzt ist: Für die 20 Tage und $\alpha = 1$ beträgt die Schranke 0.11, während $1/\lambda_j$ bis 4.9 reicht. Die unbestimmten Komponenten mit $j > p$ werden auf 0 gesetzt, weshalb der Minimierer für jede Merkmalsmatrix eindeutig ist. Abb. 1 zeigt die Auswirkung auf die gelernten Modellparameter: Die der lineare Regression schwanken von -5.8 bis 6.2 von einer Stunde zur nächsten, die der Ridge-Regression bleiben innerhalb von ± 0.7.

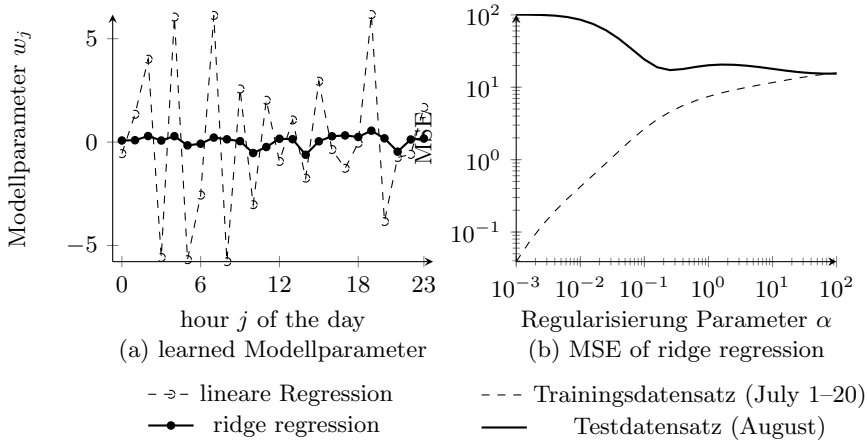

Abbildung 1: Ridge-Regression für den Messwert um 15:00 Uhr (UTC) des nächsten Tages in Krems an der Donau, gelernt aus den 24 stündlichen Temperaturen von $m = 20$ Tagen im Juli 2024. Feld (a) zeigt die 24 Modellparameter des lineares Modell, gelernt mittels lineare Regression (gestrichelte Linie, offene Kreise) und mittels Ridge-Regression mit $\alpha = 1$ (durchgezogene Linie, Punkte). Feld (b) zeigt den MSE der Ridge-Regression auf dem Trainingsdatensatz (gestrichelte Kurve) und auf den 31 Tagen des August 2024 (durchgezogene Kurve) über den Regularisierung Parameter α. Daten generiert von `pythondemos/ridgeregression.py`

Der Strafterm $\alpha\|\mathbf{w}\|_2^2$ hat eine Interpretation als eine Form von Datenerweiterung (Bishop, 1995, Gl. (10)–(11)). Insbesondere ist das Hinzufügen des Strafterm zum durchschnittlichen squared error loss der Hypothese $h^{(\mathbf{w})}$ äquivalent dazu, ihren durchschnittlichen squared error loss auf einem augmentierten Trainingsdatensatz zu berechnen (siehe Abb. 2). Dieses augmentierte Trainingsdatensatz wird aus dem ursprünglichen Trainingsdatensatz gewonnen. Für jeden Datenpunkt $\left(\mathbf{x}^{(r)}, y^{(r)}\right)$, $r = 1, \ldots, m$, wird das Merkmalsvektor $\mathbf{x}^{(r)}$ durch Realisierungen von unendlich vielen unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) $\mathbf{x}^{(r)} + \boldsymbol{\varepsilon}^{(r)}$ mit Gaußsche Zufallsvariablen $\boldsymbol{\varepsilon}^{(r)} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$ ersetzt, d. h., jede Merkmal wird unabhängig mit Varianz α gestört. Das Label $y^{(r)}$ bleibt unverändert.

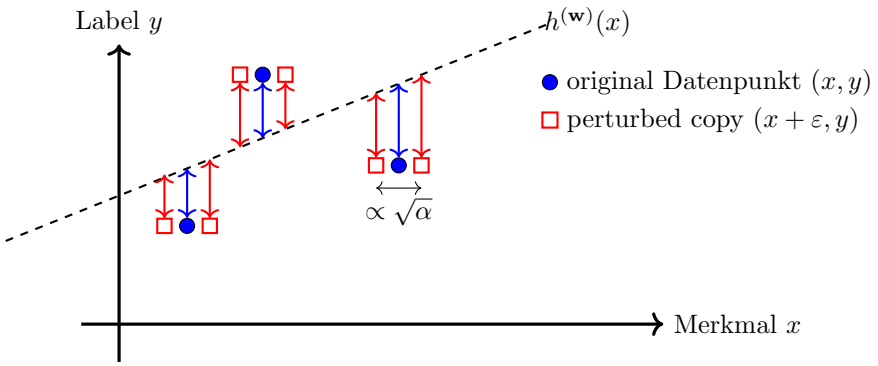

Abbildung 2: Für ein lineares Modell mit einem Merkmal ist das Hinzufügen des Strafterm $\alpha\|\mathbf{w}\|_2^2$ zum durchschnittlichen squared error loss gleichbedeutend mit der Berechnung des durchschnittlichen squared error loss auf einem erweiterten Trainingsdatensatz: jeder ursprüngliche Datenpunkt (blauer Kreis) wird durch Kopien (rote Quadrate) ersetzt, dessen Merkmal durch eine Gaußsche Zufallsvariable ε mit der Varianz α gestört wird, wobei das Label beibehalten wird. Die vertikalen Pfeile zeigen den Fehler $y - h^{(\mathbf{w})}(x)$ jedes Datenpunkt

Der Strafterm $\alpha\|\mathbf{w}\|_2^2$ ist genau der Betrag, um den der durchschnittliche squared error loss von $h^{(\mathbf{w})}$ über die unendlich vielen gestörten Kopien eines Datenpunkt seinen squared error loss auf dem ursprünglichen Datenpunkt übersteigt. Da $\mathbb{E}\{\boldsymbol{\varepsilon}^{(r)}\} = \mathbf{0}$ gilt, ist der Anstieg

$$\mathbb{E}\{(\mathbf{w}^\top \boldsymbol{\varepsilon}^{(r)})^2\} = \mathbf{w}^\top \mathbb{E}\{\boldsymbol{\varepsilon}^{(r)} (\boldsymbol{\varepsilon}^{(r)})^\top\} \mathbf{w} = \alpha\|\mathbf{w}\|_2^2.$$

Dieser Anstieg des durchschnittlichen Verlust unter Datenerweiterung ist eine Schätzung für die generalization gap (Hastie et al., 2009, Sekt. 7.4). Die Ridge-Regression, angewandt auf transformierte Merkmalsvektoren, die aus einem Kernel konstruiert werden, ist die Kernel-Ridge-Regression, die eine nichtlineare Hypothese in geschlossener Form lernt; die Störung wirkt dann auf die transformierten Merkmalsvektoren.

Synonyme: Tikhonov-Regularisierung, Gewichtsschrumpfung.

Siehe auch: regression, lineare Regression, Lasso, Regularisierung, Strafterm, SVD, condition number, Datenerweiterung, Kernel-Ridge-Regression.

Querverweise

Literatur

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

2. GeoSphere Austria (2026). GeoSphere Austria climate station archive. Da-

ta hub, station historical resource klima-v2-1d. URL: https://dataset.api.hub.geosphere.at/v1/station/hist
v2-1d.

3. Hoerl and Kennard (1970). Ridge Regression: Biased Estimation for Non-orthogonal Problems. Technometrics, 12(1), pp. 55–67.

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