Merkmalsmatrix

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 Merkmals-Matrix eines Datensatzes fasst die Merkmalsvektoren seiner Datenpunkte als ihre Zeilen zusammen, eine Zeile pro Datenpunkt und eine Spalte pro Merkmal. Ihre beiden grundlegendsten Eigenschaften sind die Anzahl der Zeilen, die der Anzahl der Datenpunkte im Datensatz entspricht, und die Anzahl der Spalten, die der Anzahl der Merkmale jedes Datenpunkts entspricht. Weitere Eigenschaften eines Datensatzes ergeben sich aus den Singulärwerten seiner Merkmals-Matrix. Die Anzahl der von null verschiedenen Singulärwerte ist der Rang, und ein Rang unterhalb der Anzahl der Merkmale führt bei der linearen Regression zu unendlich vielen Lösungen. Das Verhältnis des größten zum kleinsten von null verschiedenen Singulärwert ist die Konditionszahl der Merkmals-Matrix. Die Konditionszahl bestimmt die Anzahl der von der linearen Regression benötigten Schritte des Gradientenabstiegs sowie deren Robustheit gegenüber Störungen der Merkmale und Labels im Trainingsdatensatz.

Definition

Eine Wetterstation in Krems an der Donau zeichnet an jedem der 366 Tage des Schaltjahres 2024 acht Messwerte auf: die tiefste, höchste und mittlere Lufttemperatur, die Luftfeuchtigkeit, die Windgeschwindigkeit, den Niederschlag, die Sonnenscheindauer und den Luftdruck. Schreibt man die acht Messwerte eines Tages als eine Zeile und ordnet die Zeilen der 366 Tage untereinander an, erhält man ein zweidimensionales numerisches Array mit 366 Zeilen und acht Spalten. Dieses Array ist eine Matrix, die Merkmalsmatrix des Datensatzes, dessen Datenpunkte die 366 Tage sind, und jede ihrer Zeilen ist der Merkmalsvektor eines Tages. Die Merkmalsmatrix ist die Form, in der maschinellen Lernen-Software einen Datensatz erwartet, und Eigenschaften der auf den Datensatz angewandten ML-Methoden, etwa ob die lineare Regression eine eindeutige Lösung besitzt oder wie schnell der Gradientenabstieg konvergiert, werden durch sie bestimmt.

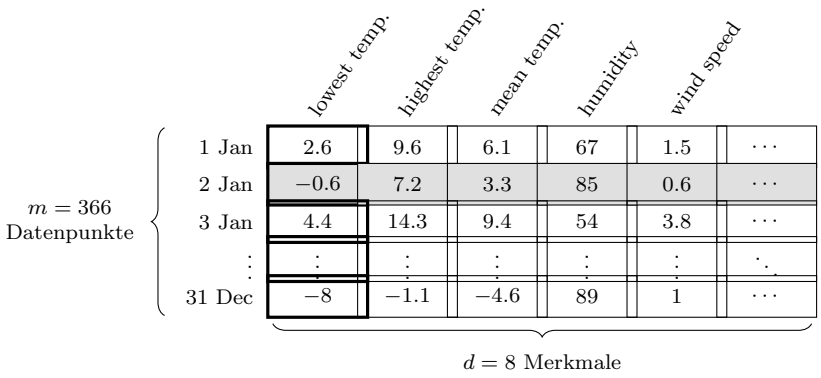


Abbildung 1: Die Merkmals-Matrix der 366 Tage des Jahres 2024 in Krems an der Donau: eine Zeile pro Tag und eine Spalte pro Messung, wobei die ersten drei Tage und der letzte sowie fünf der acht Messungen gezeigt sind (der Niederschlag, die Sonnenscheindauer und der Luftdruck sind ausgelassen). Die Temperaturen sind in Grad Celsius, die Luftfeuchtigkeit in Prozent und die Windgeschwindigkeit in Metern pro Sekunde angegeben. Die grau hinterlegte Zeile ist der transponierte Merkmalsvektor des 2. Januar; die dick umrahmte Spalte fasst die Tiefsttemperatur jedes Tages zusammen

Betrachtet sei allgemein ein Datensatz $\mathcal{D}$ aus m Datenpunkten mit Merkmalsvektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$, jeweils in $\mathbb{R}^d$. Das zeilenweise Stapeln der Merkmalsvektoren liefert die Merkmal-Matrix:

$$\mathbf{X} := \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right)^\top = \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_d^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_d^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_d^{(m)} \end{pmatrix} \in \mathbb{R}^{m \times d}.$$

Die Merkmal-Matrix hat die Größe $m \times d$, d. h., sie besitzt m Zeilen und d Spalten. Jede Zeile ist der transponierte Merkmalsvektor eines Datenpunkts, und jede Spalte enthält die Werte eines Merkmals für alle Datenpunkte (siehe Abb. 1). Viele Eigenschaften eines Datensatzes und der daraus gebildeten Lernaufgaben werden aus $\mathbf{X}$ berechnet. Die grundlegendsten sind die Anzahlen der Zeilen und Spalten, m und d.

Die Merkmal-Matrix definiert zudem eine lineare Abbildung. Ein lineares Modell mit Modellparametern $\mathbf{w}$ aus dem Parameterraum $\mathbb{R}^d$ sagt für den r-ten Datenpunkt den Wert $\mathbf{w}^\top \mathbf{x}^{(r)}$ voraus, sodass die m Vorhersagen die Einträge von $\mathbf{X}\mathbf{w}$ sind. Die Merkmal-Matrix bildet folglich jedes $\mathbf{w} \in \mathbb{R}^d$ auf den Vektor $\mathbf{X}\mathbf{w} \in \mathbb{R}^m$ ab, der mit dem Labelvektor $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(m)}\right)^\top \in \mathbb{R}^m$ verglichen wird. Die lineare Regression auf einem Trainingsdatensatz mit Merkmal-Matrix $\mathbf{X}$ und Labelvektor $\mathbf{y}$ ist die Suche nach einem Vektor $\mathbf{w}$, dessen Bild $\mathbf{X}\mathbf{w}$ in der euklidischen Norm nahe bei $\mathbf{y}$ liegt.

Die lineare Regression hat genau dann eine eindeutige Lösung, wenn $\mathbf{X}$ vollen Spalten-Rang besitzt, was $m \geq d$ erfordert (Golub and Van Loan, 2013, Sect. 5.3.1) (siehe Normalgleichungen). Ein Rang unterhalb von d führt zu unendlich vielen Parameter-Vektoren $\mathbf{w}$ mit demselben Trainingsfehler (Hastie et al., 2009, Sect. 3.2), und derjenige, den eine Methode zurückgibt, kann außerhalb des Trainingssatzes einen großen Verlust aufweisen. Die Verwendung von mehr Merkmalen als Datenpunkten (d. h. $d > m$) erzwingt diesen Fall, da $\mathbf{X}$ dann weniger Zeilen als Spalten hat, was einen vollen Spalten-Rang ausschließt.

Weitere Eigenschaften eines Datensatzes ergeben sich aus den Singulärwerten seiner Merkmal-Matrix $\mathbf{X}$. Die Singulärwertzerlegung (SVD) lässt sich schreiben als (Golub and Van Loan, 2013, Sect. 2.4)

$$\mathbf{X} = \sum_{j=1}^{\min\{m,d\}} \lambda_j \mathbf{v}^{(j)} \left(\mathbf{u}^{(j)}\right)^\top,$$

wobei die Vektoren $\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(\min\{m,d\})} \in \mathbb{R}^m$ untereinander orthonormal sind und die Vektoren $\mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(\min\{m,d\})} \in \mathbb{R}^d$ untereinander orthonormal sind. Die Singulärwerte sind geordnet als $\lambda_1 \geq \ldots \geq \lambda_{\min\{m,d\}} \geq 0$. Die Anzahl der von null verschiedenen Singulärwerte ist der Rang von $\mathbf{X}$ (Golub and Van Loan, 2013, Cor. 2.4.6). Ein Rang unterhalb von d bedeutet, dass ein Merkmal eine Linearkombination der anderen ist und über diese hinaus keine Information liefert. Ein kleiner, aber von null verschiedener Singulärwert ist die näherungsweise Form einer solchen Abhängigkeit. Gilt $m \geq d$, bilden die d rechten Singulär-Vektoren eine orthonormale Basis des $\mathbb{R}^d$, und der kleinste Singulärwert ist die kleinste euklidische Norm einer Linearkombination der Merkmal-Spalten mit Koeffizienten der Norm eins,

$$\lambda_d = \min_{\mathbf{u} \in \mathbb{R}^d,\, \|\mathbf{u}\|=1} \|\mathbf{X}\mathbf{u}\|,$$

angenommen am rechten Singulär-Vektor $\mathbf{u}^{(d)}$. Die Gleichheit gilt, da ein $\mathbf{u} = \sum_{j=1}^{d} c_j \mathbf{u}^{(j)}$ mit Norm eins auf $\|\mathbf{X}\mathbf{u}\|^2 = \sum_{j=1}^{d} c_j^2 \lambda_j^2$ führt, ein gewichtetes Mittel der quadrierten Singulärwerte. Gilt $d > m$, liefert ein $\mathbf{u}$ mit Norm eins, das orthogonal zu allen m rechten Singulär-Vektoren ist, $\mathbf{X}\mathbf{u} = \mathbf{0}$, sodass das Minimum gleich 0 ist. Ein kleiner Singulärwert besagt daher, dass eine Kombination der Merkmal-Spalten mit Koeffizienten der Norm eins nahezu der Null-Vektor ist: Die Merkmale sind nahezu kollinear, und der rechte Singulär-Vektor gibt die Kombination an. Für das Wetterjahr aus Abb. 1, bei dem jeder Messwert auf einen Stichprobenmittel von null und eine Stichprobenvarianz von eins skaliert ist, beträgt der kleinste Singulärwert 0.05 gegenüber einem größten von 36.4. Sein rechter Singulär-Vektor legt sein gesamtes Gewicht auf die tiefste, die höchste und die mittlere Temperatur, mit den Koeffizienten -0.37, -0.46 und 0.81, da die mittlere Temperatur eines Tages bis auf ein Residuum mit einer Standardabweichung von 0.03 Grad dem Mittelpunkt zwischen seiner tiefsten und seiner höchsten Temperatur entspricht. Lässt man die mittlere Temperatur weg, steigt der kleinste Singulärwert auf 4.19 (`pythondemos/featuremtx.py`).

Die Größe eines Singulärwerts hängt von den Einheiten der Spalten ab, weshalb zunächst auf eine Stichprobenvarianz von eins skaliert wird. Die Quadrate der Singulärwerte sind die Eigenwerte von $\mathbf{X}^\top\mathbf{X}$, was für Merkmale mit einem Stichprobenmittel von null dem m-Fachen ihrer Stichproben-Kovarianzmatrix entspricht.

Dieselben Eigenwerte und Eigenvektoren von $\mathbf{X}^\top\mathbf{X}$ bestimmen die Hauptkomponentenanalyse (PCA). Deren Dimensionsreduktion behält die Eigenvektoren der größten Eigenwerte als die neuen Merkmale bei, und die über die Datenpunkte gebildete Summe der quadrierten Fehler bei der Rekonstruktion der Merkmalsvektoren aus ihnen ist die Summe der weggelassenen Eigenwerte (Bishop, 2006, Sect. 12.1.2).

Eine weitere wichtige Eigenschaft einer Merkmal-Matrix ist die Konditionszahl $\kappa(\mathbf{X}) = \lambda_1/\lambda_{\mathrm{rank}(\mathbf{X})}$, das Verhältnis des größten Singulärwerts zum kleinsten von null verschiedenen. Bei vollem Spalten-Rang ist sie das Verhältnis des größten zum kleinsten Singulärwert (Golub and Van Loan, 2013, Sect. 5.3.1). Die Konditionszahl geht zweimal in die lineare Regression ein: über die Anzahl der Schritte des Gradientenabstiegs und über die Robustheit der gelernten Modellparameter gegenüber Störungen des Trainingssatzes. Die maßgebliche Größe ist in beiden Fällen ihr Quadrat $\kappa(\mathbf{X})^2$. Die Zielfunktion $f(\mathbf{w}) = (1/m)\|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$ besitzt die konstante Hesse-Matrix $(2/m)\mathbf{X}^\top\mathbf{X}$, deren Eigenwerte die mit $2/m$ skalierten quadrierten Singulärwerte von $\mathbf{X}$ sind. Die Konditionszahl von $\mathbf{X}^\top\mathbf{X}$ ist daher

$$\kappa\big(\mathbf{X}^\top\mathbf{X}\big) = \kappa(\mathbf{X})^2\,. \tag{1}$$

Das Quadrat (1) bestimmt die Konvergenz-Geschwindigkeit des Gradientenabstiegs beim empirischen Risikominimierung-Problem der linearen Regression. Ein Schritt mit exakter Liniensuche multipliziert die Differenz zwischen $f\big(\mathbf{w}^{(t)}\big)$ und dem Minimum $f\big(\widehat{\mathbf{w}}\big)$ mit höchstens $1 - 1/\kappa(\mathbf{X})^2$ (Boyd and Vandenberghe, 2004, Sect. 9.3.1). Die Anzahl der Schritte, die eine gegebene Differenz erreicht, wächst daher proportional zu $\kappa(\mathbf{X})^2$.

Abb. 2 zeigt dies für zwei Merkmal-Matrizen, die aus den Wettermessungen von Abb. 1 gebildet wurden. Die Merkmale sind die relative Luftfeuchtigkeit und die Windgeschwindigkeit, das Label ist die Höchsttemperatur, und bei jedem wurde der Stichprobenmittel abgezogen. In den aufgezeichneten Einheiten beträgt die Konditionszahl 14.5. Nach Skalierung jedes Merkmals auf eine Stichprobenvarianz von eins beträgt sie 1.45. Der Gradientenabstieg startet an dem Punkt im Abstand $\|\widehat{\mathbf{w}}\|$ von $\widehat{\mathbf{w}}$, für den er am langsamsten ist, wobei die Differenz pro Schritt genau um den Faktor in (2) schrumpft (Boyd and Vandenberghe, 2004, Sect. 9.3.2),

$$f\big(\mathbf{w}^{(t+1)}\big) - f\big(\widehat{\mathbf{w}}\big) = \Big(\frac{\kappa(\mathbf{X})^2 - 1}{\kappa(\mathbf{X})^2 + 1}\Big)^2 \Big(f\big(\mathbf{w}^{(t)}\big) - f\big(\widehat{\mathbf{w}}\big)\Big). \tag{2}$$

Von keinem Startpunkt aus schrumpft die Differenz weniger, da dieser Faktor die Schrumpfung jedes Schritts mit exakter Liniensuche bei einer quadratischen

Zielfunktion beschränkt (Bertsekas, 2016, Prop. 1.3.1). Eine Differenz von 10^{-6} der anfänglichen erfordert dann 731 Schritte in den aufgezeichneten Einheiten und 7 nach der Skalierung.

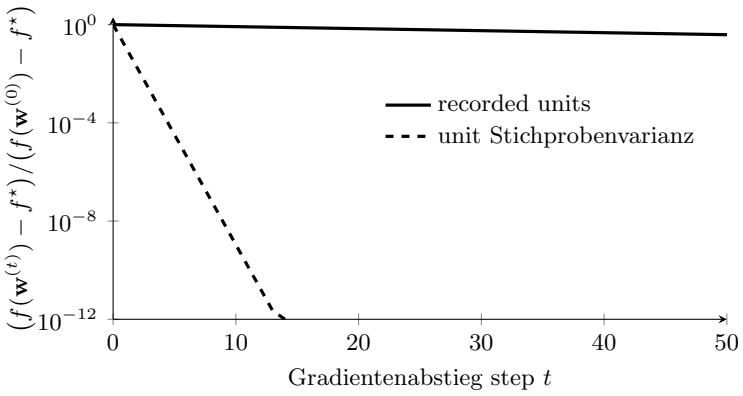


Abbildung 2: Die Konvergenz des Gradientenabstiegs mit exakter Liniensuche für die Merkmale relative Luftfeuchtigkeit und Windgeschwindigkeit der 366 Tage, in ihren aufgezeichneten Einheiten (durchgezogene Kurve) und nach Skalierung jedes Merkmals auf eine Stichprobenvarianz von eins (gestrichelte Kurve). Gezeigt ist die Differenz zwischen $f\big(\mathbf{w}^{(t)}\big)$ und dem Minimum $f^\star = \min_{\mathbf{w}\in\mathbb{R}^2} f(\mathbf{w})$, relativ zu ihrem Anfangswert, über die ersten 50 Schritte. Beide Läufe starten an dem Punkt im Abstand $\|\widehat{\mathbf{w}}\|$ von $\widehat{\mathbf{w}}$, für den der Gradientenabstieg am langsamsten ist; die gestrichelte Kurve endet dort, wo die Differenz unter 10^{-12} fällt. Daten erzeugt durch `pythondemos/featuremtx.py`

Die Konditionszahl $\kappa(\mathbf{X})$ bestimmt auch die Robustheit der von der linearen Regression gelieferten Modellparameter $\widehat{\mathbf{w}}$ gegenüber Störungen der Trainingsdatensatz. Sei $\mathbf{X}$ von vollem Spalten-Rang. Die lineare Regression liefert die Modellparameter $\widehat{\mathbf{w}} = \mathbf{X}^+\mathbf{y}$ mit der Pseudoinversen $\mathbf{X}^+ = \big(\mathbf{X}^\top\mathbf{X}\big)^{-1}\mathbf{X}^\top$ (Golub and Van Loan, 2013, Sect. 5.5.2). Sie sind die eindeutige Lösung der Normalgleichungen $\mathbf{X}^\top\mathbf{X}\widehat{\mathbf{w}} = \mathbf{X}^\top\mathbf{y}$, die besagen, dass der Gradient von f bei $\widehat{\mathbf{w}}$ verschwindet.

Betrachtet werden Störungen der Trainingsdatensatz, die $\mathbf{X}$ zu $\mathbf{X}+\delta\mathbf{X}$ und $\mathbf{y}$ zu $\mathbf{y}+\delta\mathbf{y}$ ändern, wobei $\mathbf{X}+\delta\mathbf{X}$ weiterhin vollen Spalten-Rang hat, und es seien $\widehat{\mathbf{w}}+\delta\widehat{\mathbf{w}}$ die durch lineare Regression auf der gestörten Trainingsdatensatz gelernten Modellparameter. Sie lösen die Normalgleichungen der gestörten Trainingsdatensatz,

$$(\mathbf{X}+\delta\mathbf{X})^\top(\mathbf{X}+\delta\mathbf{X})\big(\widehat{\mathbf{w}}+\delta\widehat{\mathbf{w}}\big) = (\mathbf{X}+\delta\mathbf{X})^\top(\mathbf{y}+\delta\mathbf{y}).$$

Bringt man die Terme mit $\widehat{\mathbf{w}}$ auf die rechte Seite und verwendet $\mathbf{X}^\top\mathbf{r} = \mathbf{0}$ für das Residuum der Methode der kleinsten Quadrate $\mathbf{r} = \mathbf{y} - \mathbf{X}\widehat{\mathbf{w}}$, was den

Normalgleichungen der ungestörten Trainingsdatensatz entspricht, erhält man die exakte Gleichung

$$(\mathbf{X}+\delta\mathbf{X})^\top(\mathbf{X}+\delta\mathbf{X})\,\delta\widehat{\mathbf{w}} = (\mathbf{X}+\delta\mathbf{X})^\top\big(\delta\mathbf{y}-\delta\mathbf{X}\,\widehat{\mathbf{w}}\big)+\delta\mathbf{X}^\top\mathbf{r}.$$

Das Auflösen nach $\delta\widehat{\mathbf{w}}$ wendet die Pseudoinverse von $\mathbf{X}+\delta\mathbf{X}$ auf das Rauschen in den Labels und auf das Rauschen in den Merkmale, das auf $\widehat{\mathbf{w}}$ wirkt, an, sowie die Inverse von $(\mathbf{X}+\delta\mathbf{X})^\top(\mathbf{X}+\delta\mathbf{X})$ auf das Rauschen in den Merkmale, das auf das Residuum wirkt.

Der größte Singulärwert von $\mathbf{X}^+$ ist $1/\lambda_d$, und derjenige von $\big(\mathbf{X}^\top\mathbf{X}\big)^{-1}$ ist $1/\lambda_d^2$ (Golub and Van Loan, 2013, Sect. 5.3.6). Dasselbe gilt für $\mathbf{X}+\delta\mathbf{X}$ mit seinem eigenen kleinsten Singulärwert. Sei der größte Singulärwert von $\delta\mathbf{X}$ höchstens $\varepsilon\lambda_1$ und die euklidische Norm von $\delta\mathbf{y}$ höchstens $\varepsilon\|\mathbf{y}\|$, mit $\varepsilon\,\kappa(\mathbf{X})<1$. Der kleinste Singulärwert von $\mathbf{X}+\delta\mathbf{X}$ ist dann mindestens $\lambda_d-\varepsilon\lambda_1$ (Golub and Van Loan, 2013, Cor. 8.6.2), und das Anwenden euklidischer Normen in der exakten Gleichung liefert die Störungsschranke

$$\|\delta\widehat{\mathbf{w}}\| \le \frac{\varepsilon\,\kappa(\mathbf{X})}{1-\varepsilon\,\kappa(\mathbf{X})}\Big(\frac{\|\mathbf{y}\|}{\lambda_1}+\|\widehat{\mathbf{w}}\|\Big)+\frac{\varepsilon\,\kappa(\mathbf{X})^2}{\big(1-\varepsilon\,\kappa(\mathbf{X})\big)^2}\,\frac{\|\mathbf{r}\|}{\lambda_1}. \tag{3}$$

Die Schranke (3) gilt für jede Störung mit $\varepsilon\,\kappa(\mathbf{X})<1$.

Die Schranke (3) überträgt sich auf die Vorhersage für jeden gegebenen Merkmalsvektor $\mathbf{x}$, wie etwa den eines Test-Datenpunkts. Das gelernte lineare Modell sagt vor der Störung $\mathbf{x}^\top\widehat{\mathbf{w}}$ und danach $\mathbf{x}^\top\big(\widehat{\mathbf{w}}+\delta\widehat{\mathbf{w}}\big)$ vorher. Die Vorhersage ändert sich daher um $\mathbf{x}^\top\delta\widehat{\mathbf{w}}$, und die Cauchy-Schwarz-Ungleichung beschränkt diese Änderung durch

$$\big|\mathbf{x}^\top\delta\widehat{\mathbf{w}}\big| \le \|\mathbf{x}\|\,\|\delta\widehat{\mathbf{w}}\| \le \|\mathbf{x}\|\left(\frac{\varepsilon\,\kappa(\mathbf{X})}{1-\varepsilon\,\kappa(\mathbf{X})}\Big(\frac{\|\mathbf{y}\|}{\lambda_1}+\|\widehat{\mathbf{w}}\|\Big)+\frac{\varepsilon\,\kappa(\mathbf{X})^2}{\big(1-\varepsilon\,\kappa(\mathbf{X})\big)^2}\,\frac{\|\mathbf{r}\|}{\lambda_1}\right). \tag{4}$$

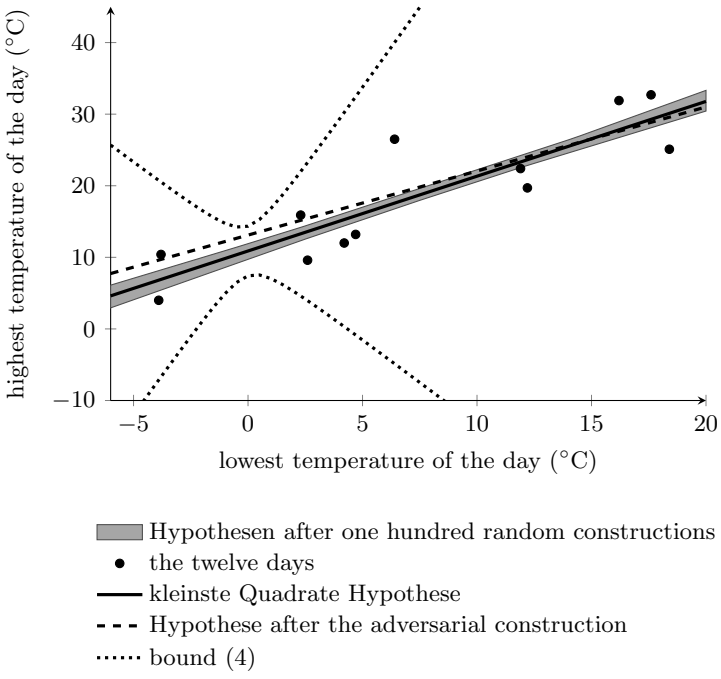
Die Änderung der Vorhersage beträgt daher höchstens das $\|\mathbf{x}\|$-Fache der rechten Seite der Schranke (3) und wächst mit der Norm des Merkmalsvektors.

Abb. 3 zeigt die Robustheit von $\widehat{\mathbf{w}}$ gegenüber Störungen auf einem verwandten, aber anderen und kleineren Datensatz, den ersten Tagen der zwölf Monate des Jahres 2024, zwölf der Zeilen aus Abb. 1. Jeder Tag hat den Merkmalsvektor $\mathbf{x}=(1,x)^\top$, wobei x seine Tiefsttemperatur ist, und seine Höchsttemperatur als Label. Eine Hypothese des linearen Modells hat dann den Achsenabschnitt w_1 und die Steigung w_2, und die Merkmal-Matrix hat eine Spalte aus Einsen und die Spalte der Tiefsttemperaturen. Ihre Konditionszahl beträgt 14.9, und das Residuum der Methode der kleinsten Quadrate beträgt das 0.20-Fache der Norm von $\mathbf{y}$.

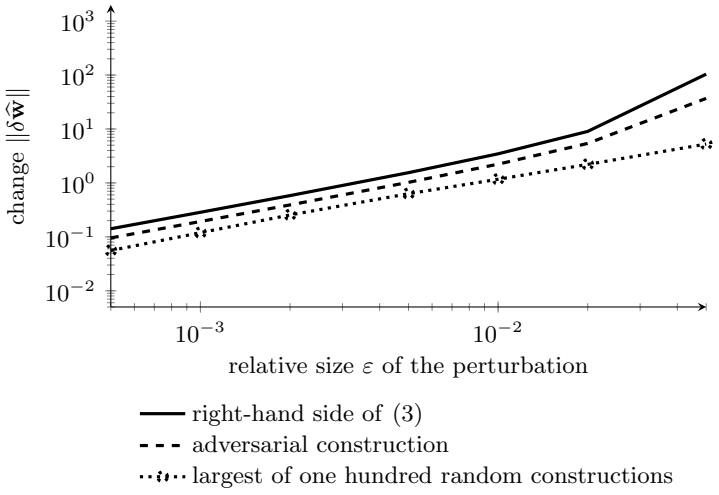
Zwei Konstruktionen einer Störung der relativen Größe $\varepsilon=0.01$ werden verglichen. Die erste wird adversariell gewählt, um die Änderung der durch lineare Regression gelernten Modellparameter so groß wie möglich zu machen: $\delta\mathbf{X}=-\varepsilon\lambda_1\mathbf{v}^{(2)}\big(\mathbf{u}^{(2)}\big)^\top$ und $\delta\mathbf{y}=\pm\varepsilon\|\mathbf{y}\|\mathbf{v}^{(2)}$, mit demjenigen Vorzeichen, für das sich $\widehat{\mathbf{w}}$ stärker bewegt. Sie ändert $\widehat{\mathbf{w}}$ um die Norm 2.23, was 64 Prozent der

rechten Seite der Schranke (3) entspricht. Die zweite Konstruktion verwendet 100 Realisierungen von zufälligen $\delta\mathbf{X}$ und $\delta\mathbf{y}$. Sie werden mit unabhängig und identisch verteilt (u.i.v.) gaußschen Zufallsvariablen (ZV) als Einträgen erzeugt und dann so normiert, dass der größte Singulärwert von $\delta\mathbf{X}$ gleich $\varepsilon\lambda_1$ und die euklidische Norm von $\delta\mathbf{y}$ gleich $\varepsilon\|\mathbf{y}\|$ ist. Die größte Änderung unter den 100 Realisierungen beträgt 1.17, was 34 Prozent der rechten Seite der Schranke (3) entspricht.

Für einen Tag mit der Tiefsttemperatur x hat der Merkmalsvektor $(1, x)^\top$ die Norm $\sqrt{1+x^2}$. Nach der Vorhersageschranke (4) ändert sich die Vorhersage daher um höchstens das $\sqrt{1+x^2}$-Fache der rechten Seite der Schranke (3): 3.5 Grad bei $x = 0$ und 69 Grad bei $x = 20$. Die beiden gepunkteten Kurven in Abb. 3(a) zeichnen diese Schranke nach und verlassen den dargestellten Bereich. Die aus den zufälligen Störungen gelernten Hypothesen bleiben über den dargestellten Bereich innerhalb von 1.6 Grad von der gelernten Hypothese, und die Hypothese nach der adversariellen Konstruktion innerhalb von 3.1 Grad. Teilabbildung (b) wiederholt den Vergleich für ε von $5 \cdot 10^{-4}$ bis $5 \cdot 10^{-2}$. Die adversarielle Konstruktion erreicht bis $\varepsilon = 0.02$ mehr als die Hälfte der rechten Seite der Schranke (3), die zufällige Konstruktion bei jedem ε weniger als die Hälfte.



(a)



(b)

Abbildung 3: Die Störungsschranke für die Merkmale $(1, x)$, wobei x die tiefste Temperatur eines Tages und die höchste Temperatur das Label ist, über die ersten Tage der zwölf Monate des Jahres 2024 (ausgefüllte Kreise). (a) Die kleinste Quadrate-Hypothese (durchgezogene Linie), die nach der adversariellen Konstruktion einer Störung der relativen Größe $\varepsilon = 0.01$ gelernte Hypothese (gestrichelte Linie), der Bereich der nach 100 zufälligen Konstruktionen derselben relativen Größe gelernten Hypothesen (graues Band) und die Schranke (4) oberhalb und unterhalb der gelernten Hypothese (gepunktete Kurven, die den dargestellten Bereich verlassen). (b) Die Änderung der gelernten Modellparameter in Abhängigkeit von ε: die rechte Seite der Schranke (3) (durchgezogene Kurve), die adversarielle Konstruktion (gestrichelte Kurve) und die größte Änderung unter 100 zufälligen Konstruktionen (gepunktete Kurve mit Kreisen). Daten erzeugt durch `pythondemos/featuremtx.py`

Synonyme: Designmatrix, Datenmatrix.

Siehe auch: Datensatz, Datenpunkt, Merkmalsvektor, Merkmal, Array, Matrix, Designmatrix, lineares Modell, Parameterraum, SVD, Singulärwert, Rang, Konditionszahl, Stichproben-Kovarianzmatrix, PCA, Dimensionsreduktion, Spaltenraum.

Querverweise

Literatur

1. Bertsekas (2016). Nonlinear Programming. 3rd ed., Athena Scientific, Belmont, MA, USA.

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

3. Boyd and Vandenberghe (2004). Convex Optimization. Cambridge Univ. Press, Cambridge, U.K.

4. Golub and Van Loan (2013). Matrix Computations. 4th ed., The Johns Hopkins Univ. Press, Baltimore, MD, USA.

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