Norm

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 Norm auf einem Vektorraum ist eine Funktion, die jedem Vektor eine nichtnegative Zahl, seine Länge, zuordnet. Eine Norm ist definit und homogen und erfüllt die Dreiecksungleichung. Jede Norm definiert eine Metrik über die Norm der Differenz zweier Vektoren, und jedes Skalarprodukt induziert eine Norm, aber nicht jede Norm wird von einem solchen induziert. Die ℓ_p-Normen auf $\mathbb{R}^d$ umfassen die ℓ_1-Norm, die Euclidean norm und die ℓ_∞-Norm; unter ihnen wird nur die Euclidean norm von einem Skalarprodukt induziert. In der machine learning (ML) werden Normen zur Konstruktion von Verlustfunktionen und Regularisierungsterme verwendet.

Definition

Die stündlichen Temperaturaufzeichnungen eines Tages an einer Wetterstation bilden die Merkmalsvektor $\mathbf{x} \in \mathbb{R}^d$ dieses Tages, mit einem Eintrag pro Stunde. Viele machine learning (ML)-Methoden verwenden ein Maß für die Größe eines solchen Vektor oder für die Abstand zwischen zweien davon. Eine Norm liefert beides: Sie ordnet jedem Vektor eine Länge zu, und die Länge der Differenz zweier Vektoren ist deren Abstand.

Formal ist eine Norm $\|\cdot\|$ auf einem Vektorraum $\mathcal{V}$ über dem field $\mathbb{R}$ eine Funktion $\|\cdot\| : \mathcal{V} \to \mathbb{R}_+$, die die folgenden Bedingungen erfüllt (Horn and Johnson, 2013). Diese Bedingungen gelten für alle Vektoren $\mathbf{u}, \mathbf{v} \in \mathcal{V}$ und alle Skalare $\beta \in \mathbb{R}$, wobei $|\beta|$ der Absolutwert von β ist:

- Definitheit: $\|\mathbf{u}\| = 0 \iff \mathbf{u} = \mathbf{0}$;
- Homogenität: $\|\beta\mathbf{u}\| = |\beta| \cdot \|\mathbf{u}\|$;
- Dreiecksungleichung: $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

Die gleichen Bedingungen definieren eine Norm auf einem Vektorraum über dem field der komplexen Zahlen $\mathbb{C}$, mit $\beta \in \mathbb{C}$ und $|\beta|$ als dessen Betrag. Eine Funktion, die Homogenität und die Dreiecksungleichung erfüllt, aber für einige von null verschiedene Vektor den Wert null annimmt, ist eine Seminorm, keine Norm: Auf $\mathbb{R}^2$ ist die Abbildung $\mathbf{u} = (u_1, u_2)^\top \mapsto |u_1|$, die nur den ersten Eintrag berücksichtigt, eine Seminorm, die auf $(0, 1)^\top$ null ist.

Jede Norm definiert eine Metrik durch $d(\mathbf{u}, \mathbf{v}) := \|\mathbf{u} - \mathbf{v}\|$, sodass $(\mathcal{V}, d(\cdot, \cdot))$ ein metric space ist. Manche Normen werden von einem Skalarprodukt $\langle \cdot, \cdot \rangle$ induziert, einer Funktion zweier Vektoren, die symmetrisch, in jedem Argument linear und positiv definit ist (auf $\mathbb{R}^d$ gilt $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v}$). Auf einem Skalarprodukt-Raum ist $\|\mathbf{u}\| := \sqrt{\langle \mathbf{u}, \mathbf{u} \rangle}$ eine Norm (Axler, 2024, Thm. 6.14 und Thm. 6.17), und die Norm eines Hilbertraum ist von dieser Form. Eine induzierte Norm erfüllt die Parallelogrammgleichung $\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$, die durch Ausmultiplizieren von $\langle \mathbf{u} \pm \mathbf{v}, \mathbf{u} \pm \mathbf{v} \rangle$ entsteht.

Für den Vektorraum $\mathcal{V} = \mathbb{R}^d$ der Merkmalsvektoren $\mathbf{x} = (x_1, \ldots, x_d)^\top$ und eine Zahl $p \geq 1$ ist die ℓ_p-Norm von $\mathbf{x}$ definiert als $\|\mathbf{x}\|_p = \left(\sum_{j=1}^{d} |x_j|^p \right)^{1/p}$. Sie erfüllt die Dreiecksungleichung nach der Minkowski-Ungleichung und ist daher eine Norm (Horn and Johnson, 2013, Sekt. 5.2, Gl. (5.2.4)). Zu ihren Ausprägungen gehören die ℓ_1-Norm $\|\mathbf{x}\|_1 = \sum_{j=1}^{d} |x_j|$ und die ℓ_2-Norm $\|\mathbf{x}\|_2 = \left(\sum_{j=1}^{d} x_j^2 \right)^{1/2}$ (die Euclidean norm); der Grenzwert $p \to \infty$ ergibt die ℓ_∞-Norm $\|\mathbf{x}\|_\infty = \max_{j=1,\ldots,d} |x_j|$. Die ℓ_2-Norm wird durch das Skalarprodukt $\mathbf{u}^\top \mathbf{v}$ induziert und ist die einzige ℓ_p-Norm, die von einem Skalarprodukt induziert wird. Für $\mathbf{u} = (1, 0)^\top$ und $\mathbf{v} = (0, 1)^\top$ in $\mathbb{R}^2$ haben sowohl $\mathbf{u} + \mathbf{v}$ als auch $\mathbf{u} - \mathbf{v}$ die ℓ_p-Norm $2^{1/p}$, sodass die linke Seite der Parallelogrammgleichung $2 \cdot 2^{2/p}$ ist, während die rechte Seite 4 beträgt, und beide stimmen nur für $p = 2$ überein. Die linke Seite ergibt 8 für die ℓ_1-Norm und 2 für die ℓ_∞-Norm. Die drei Normen unterscheiden sich darin, welche Vektoren sie als von Einheitslänge ansehen. Die Einheitskugel $\mathbb{S}^{(1)} := \left\{ \mathbf{u} \in \mathbb{R}^2 : \|\mathbf{u}\| = 1 \right\}$ einer Norm auf $\mathbb{R}^2$ enthält diese Vektoren, und Abb. 1 zeigt sie für die ℓ_1-, die ℓ_2- und die ℓ_∞-Norm.

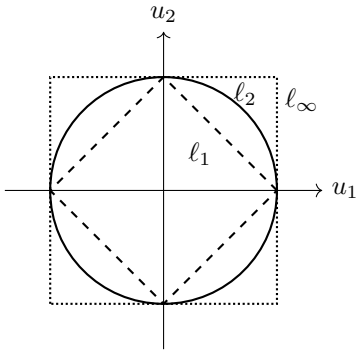

Abbildung 1: Einheitskugeln $\mathbb{S}^{(1)}$ der ℓ_p-Normen auf $\mathbb{R}^2$, für Vektoren $\mathbf{u} = (u_1, u_2)^\top$ und $p = 1$ (gestrichelte Raute), $p = 2$ (durchgezogener Kreis) und $p = \infty$ (gepunktetes Quadrat)

ML-Methoden verwenden Normen zur Konstruktion von Verlustfunktionen und Regularisierungsterme. In der lineare Regression stapelt die label vector $\mathbf{y} \in \mathbb{R}^m$ die Labels eines Trainingsdatensatz der Größe m und die Merkmalsma-

trix $\mathbf{X} \in \mathbb{R}^{m \times d}$ deren Merkmalsvektoren. Der durchschnittliche squared error loss einer linearen Hypothese mit Modellparameter $\mathbf{w}$ ist dann die skalierte quadrierte Euclidean norm $(1/m)\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ des Residuums $\mathbf{y} - \mathbf{X}\mathbf{w}$. Ridge-Regression fügt zu diesem Trainingsfehler den Strafterm $\alpha\|\mathbf{w}\|_2^2$ mit einem Hyperparameter $\alpha > 0$ hinzu, während das least absolute shrinkage and selection operator (Lasso) $\alpha\|\mathbf{w}\|_1$ addiert (Tibshirani, 1996).

Die Wahl der Norm bestimmt die gelernte Hypothese durch empirische Risikominimierung (ERM). Betrachtet werden die beiden Optimierungsprobleme

$$\min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2 \quad \text{und} \quad \min_{\mathbf{w} \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_1,$$

die sich nur in der Norm des Residuums unterscheiden. Das Quadrieren der ℓ_2-Norm ändert die Menge der Minimierer nicht, sodass das Minimieren der ℓ_2-Norm des Residuums äquivalent zum Minimieren des durchschnittlichen squared error loss ist. Der Unterschied zwischen induzierten und anderen Normen ist hier relevant. Eine induzierte Norm definiert Winkel und orthogonale Projektionen, und die quadrierte ℓ_2-Norm des Residuums ist eine quadratische Funktion von $\mathbf{w}$, sodass das least squares-Problem eine geschlossene Lösung besitzt, die aus den Normalgleichungen gewonnen wird (Boyd and Vandenberghe, 2004, Sekt. 1.2). Die ℓ_2-Norm ist daher die Norm der Verlustfunktion der lineare Regression. Das Minimieren der ℓ_1-Norm des Residuums ist ein lineares Programm (Boyd and Vandenberghe, 2004, Sekt. 6.1.1), das keine geschlossene Lösung besitzt und iterativ gelöst wird, ebenso wie das Lasso-Problem mit seiner ℓ_1-Norm-Strafe.

Abb. 2 zeigt ein Trainingsdatensatz aus vier Datenpunkte, von denen einer ein Ausreißer ist, und zwei lineare Hypothesen $h(x) = wx$, die daraus gelernt wurden. Das Minimieren der quadrierten ℓ_2-Norm des Residuums ergibt die Steigung $\widehat{w} = 46/30$. Diese Lösung wird zum Ausreißer hin verschoben, da das Quadrieren das größte Residuum zum Hauptanteil des Trainingsfehler macht. Das Minimieren der ℓ_1-Norm ergibt $\widehat{w} = 1$, was die drei verbleibenden Datenpunkte exakt vorhersagt und ein einzelnes großes Residuum beim Ausreißer belässt. Die mit der ℓ_1-Norm gelernte Hypothese ist robust gegenüber dem Ausreißer in dem Sinne, dass ein ungewöhnlicher Datenpunkt sie kaum verschiebt (Boyd and Vandenberghe, 2004, Sekt. 6.1.2), während die durch Minimieren der quadrierten ℓ_2-Norm gelernte nicht robust ist (siehe robustness). Beide Steigungen werden aus dem Trainingsdatensatz gelernt; welche der beiden Hypothesen das geringere Risiko auf außerhalb liegenden Datenpunkte hat, hängt davon ab, ob dort ebenfalls Ausreißer vorkommen.

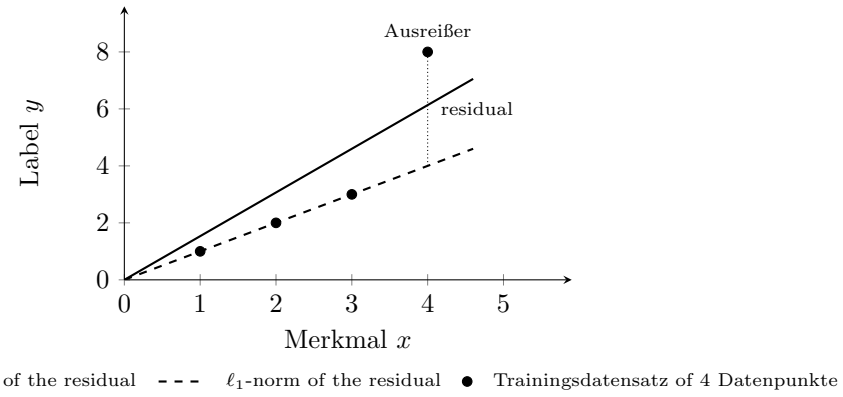

Abbildung 2: Zwei lineare Hypothesen $h(x) = wx$, gelernt aus demselben Trainingsdatensatz von vier Datenpunkte (ausgefüllte Kreise) durch Minimieren einer Norm des Residuums. Die quadrierte ℓ_2-Norm ergibt die durchgezogene Linie mit Steigung $\widehat{w} = 46/30$, verschoben zum Ausreißer; die ℓ_1-Norm ergibt die gestrichelte Linie mit Steigung $\widehat{w} = 1$, die die drei verbleibenden Datenpunkte exakt vorhersagt und ein großes Residuum (gepunktetes vertikales Segment) beim Ausreißer belässt. Daten erzeugt mit `pythondemos/norm.py`

Siehe auch: Vektorraum, Metrik, metric space, Euclidean space, Hilbertraum, Skalarprodukt, Euclidean norm, Regularisierungsterm, lineare Regression, Lasso.

Querverweise

Literatur

1. Axler (2024). Linear Algebra Done Right. 4th ed., Springer, Cham, Switzerland.

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

3. Horn and Johnson (2013). Matrix Analysis. 2nd ed., Cambridge Univ. Press, New York, NY, USA.

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