Gradient

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

Der Gradient einer reellwertigen Funktion ist ein Vektor, der die lokale lineare Approximation der Funktion bestimmt. Die Einträge des Gradienten sind die partielle Ableitungen der Funktion. Ihre Existenz allein macht die Funktion jedoch nicht differentiable: Der von ihnen gebildete Vektor muss keine lokale lineare Approximation liefern. Für eine konvex Funktion genügt die Existenz aller partielle Ableitungen an einem Punkt bereits, damit sie dort differentiable ist. Geometrisch steht ein von Null verschiedener Gradient orthogonal auf den Niveaumengen und zeigt in die Richtung des steilsten Anstiegs. In der machine learning (ML) bestimmen Gradienten der empirische Risikominimierung (ERM) Zielfunktion das Update von Gradientenabstieg-Verfahren und werden für tiefe Netze durch Backpropagation berechnet.

Definition

Ein lineares Modell für die maximale Temperatur in Krems an der Donau besitzt zwei Modellparameter, und jede Änderung eines dieser Parameter verändert den mittleren squared error loss auf dem Trainingsdatensatz. Das Training eines machine learning (ML) Modell passt seine Modellparameter $\mathbf{w}$ an, um eine Verlust $f(\mathbf{w})$ zu verringern, wozu bekannt sein muss, wie sich f ändert, wenn sich $\mathbf{w}$ bewegt. Der Gradient einer reellwertigen Funktion $f : \mathbb{R}^d \to \mathbb{R} : \mathbf{w} \mapsto f(\mathbf{w})$ beantwortet dies, indem er eine lokale lineare Approximation von f bestimmt. Formal ist der Gradient von f an einem Punkt $\mathbf{w}' \in \mathbb{R}^d$ ein Vektor $\mathbf{g} \in \mathbb{R}^d$, so dass

$$\lim_{\mathbf{w} \to \mathbf{w}'} \frac{f(\mathbf{w}) - \big(f(\mathbf{w}') + \mathbf{g}^\top (\mathbf{w} - \mathbf{w}')\big)}{\|\mathbf{w} - \mathbf{w}'\|_2} = 0. \tag{1}$$

Existiert ein solcher Vektor, so ist er eindeutig, die Funktion f ist an $\mathbf{w}'$ differentiable, und der Vektor wird mit $\nabla f(\mathbf{w}')$ oder $\nabla f(\mathbf{w})\big|_{\mathbf{w}'}$ bezeichnet (Rudin, 1976, Def. 9.11 und Thm. 9.12). Die Funktion $f(\mathbf{w}') + \big(\nabla f(\mathbf{w}')\big)^\top (\mathbf{w} - \mathbf{w}')$ ist die lokale lineare Approximation von f an $\mathbf{w}'$ (siehe Abb. 1).

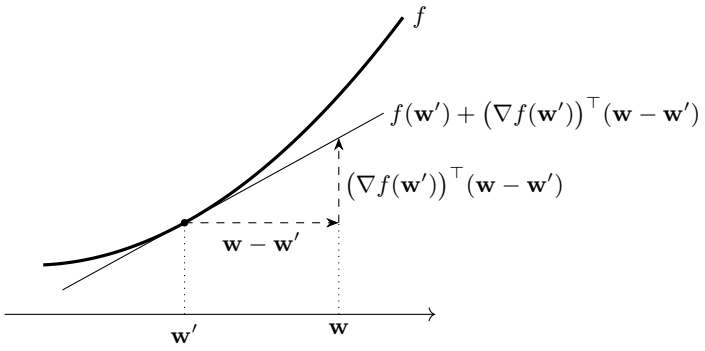

Abbildung 1: Eine differentiable Funktion $f : \mathbb{R} \to \mathbb{R}$ (fette Kurve) und ihre lokale lineare Approximation an einem Punkt $\mathbf{w}'$ (dünne Gerade). Das gestrichelte Steigungsdreieck zeigt die Verschiebung $\mathbf{w} - \mathbf{w}'$ (horizontale Kathete) und die daraus resultierende Änderung $\big(\nabla f(\mathbf{w}')\big)^\top (\mathbf{w} - \mathbf{w}')$ (vertikale Kathete) der lokalen linearen Approximation

Die Einträge des Gradienten sind die partielle Ableitungen von f, $\nabla f(\mathbf{w}) = \big(\partial f/\partial w_1, \ldots, \partial f/\partial w_d\big)^\top$ (Rudin, 1976, Thm. 9.17); die partielle Ableitung $\partial f/\partial w_j$ ist die Änderungsrate von f, wenn nur der j-te Eintrag von $\mathbf{w}$ variiert. Die Existenz aller partielle Ableitungen von f an einem Punkt garantiert nicht, dass dort auch der Gradient existiert (Rudin, 1976, Sekt. 9.16 und Thm. 9.21). Zusätzliche Annahmen an f stellen die Implikation wieder her; Konvexität ist eine hinreichende Bedingung. Wenn eine konvex Funktion $f : \mathbb{R}^d \to \mathbb{R}$ an einem Punkt $\mathbf{w}'$ partielle Ableitungen $\partial f/\partial w_j$ für $j = 1, \ldots, d$ besitzt, dann ist f an $\mathbf{w}'$ differentiable, und der Gradient $\nabla f(\mathbf{w}')$ ist der Vektor dieser partielle Ableitungen (Rockafellar, 1970, Thm. 25.2).

An jedem Punkt, an dem f differentiable ist und $\nabla f \neq \mathbf{0}$ gilt, steht der Gradient orthogonal auf der Niveaumenge von f durch diesen Punkt, also der Menge $\{\mathbf{w} : f(\mathbf{w}) = c\}$ der Punkte, an denen f denselben Wert c annimmt. Er zeigt außerdem in die Richtung des steilsten Anstiegs von f. Der negative Gradient $-\nabla f(\mathbf{w}')$ zeigt in die Richtung des steilsten Abstiegs, und Gradientenabstieg schreitet wiederholt in diese Richtung (siehe Abb. 2; Boyd and Vandenberghe, 2004, Sekt. 9.4.1). An einem lokalen minimum einer differentiable Funktion $f : \mathbb{R}^d \to \mathbb{R}$ kann es keine Abstiegsrichtung geben, und folglich muss der Gradient verschwinden (Nesterov, 2004, Thm. 1.2.1) (siehe zero-gradient condition).

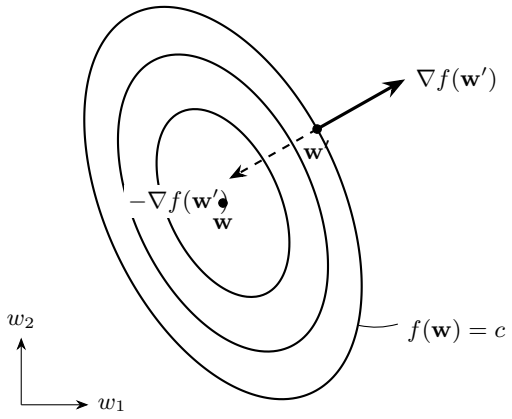

Abbildung 2: Niveaumengen $f(\mathbf{w}) = c$ (die Kurven, entlang derer f eine Konstante c annimmt) einer differentiable Funktion f von zwei Variablen w_1, w_2 (Achsen, unten links), mit einem minimum bei $\widehat{\mathbf{w}}$. Der Gradient $\nabla f(\mathbf{w}')$ (durchgezogener Pfeil) steht orthogonal auf der Niveaumenge durch $\mathbf{w}'$ (die Kurve durch $\mathbf{w}'$) und zeigt in die Richtung des steilsten Anstiegs. Der negative Gradient $-\nabla f(\mathbf{w}')$ (gestrichelter Pfeil) zeigt in die Richtung des steilsten Abstiegs, also zu kleineren Werten von f. Daten erzeugt mit `pythondemos/gradient.py`

Gradienten werden beim Training von ML Modelle verwendet. Sie bestimmen das Update der Modellparameter, das die auf einem Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \left\{\left(\mathbf{x}^{(r)}, y^{(r)}\right)\right\}_{r=1}^{m}$ entstehende Verlust minimiert, was das Ziel der empirische Risikominimierung (ERM) ist. Beispielsweise minimiert die lineare Regression die durch

$$f(\mathbf{w}) := \frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right)^2.$$

definierte Zielfunktion. Diese Funktion ist konvex und differentiable, mit dem Gradienten

$$\nabla f(\mathbf{w}) = -\frac{2}{m} \sum_{r=1}^{m} \left(y^{(r)} - \mathbf{w}^\top \mathbf{x}^{(r)}\right) \mathbf{x}^{(r)}. \tag{2}$$

Das Nullsetzen des Gradienten (2) ergibt die Normalgleichungen der lineare Regression. Für ein tiefes künstliches neuronales Netz (KNN), dessen Aktivierungsfunktionen, Verlust und alle weiteren Operationen, aus denen es zusammengesetzt ist, differentiable sind, ist auch die ERM Zielfunktion nach der Kettenregel (Rudin, 1976, Thm. 9.15) differentiable. Der Gradient dieser Zielfunktion kann durch Backpropagation berechnet werden (Goodfellow et al., 2016, Sekt. 6.5; Rumelhart et al., 1986).

Die Definition überträgt sich auf eine reellwertige Funktion $f : \mathcal{H} \to \mathbb{R}$ auf einem reellen Hilbertraum $\mathcal{H}$, einem Hilbertraum über $\mathbb{R}$. Im definierenden

Limes (1) ersetzt das Skalarprodukt $\langle \mathbf{g}, \mathbf{w}-\mathbf{w}' \rangle$ von $\mathcal{H}$ den Term $\mathbf{g}^\top(\mathbf{w}-\mathbf{w}')$. Die Norm von $\mathcal{H}$ ersetzt die Euclidean norm $\|\mathbf{w}-\mathbf{w}'\|_2$ (Bauschke and Combettes, 2011, Remark 2.44 und Def. 2.45).

Siehe auch: Funktion, Vektor, differentiable, partielle Ableitung, Gradientenabstieg, zero-gradient condition, konvex, Hilbertraum.

Querverweise

Literatur

1. Bauschke and Combettes (2011). Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer Science+Business Media, New York, NY, USA.

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

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

4. Rockafellar (1970). Convex Analysis. Princeton Univ. Press, Princeton, NJ, USA.

5. Rudin (1976). Principles of Mathematical Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

6. Rumelhart et al. (1986). Learning Representations by Back-Propagating Errors. Nature, 323, pp. 533–536.

7. Nesterov (2004). Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic, Boston, MA, USA.