konvex

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 Teilmenge des euklidischen Raum ist konvex, wenn sie das Streckensegment zwischen beliebigen zwei ihrer Punkte enthält. Eine Funktion ist konvex, wenn ihr Epigraph eine konvexe Menge ist. Ein konvexes Optimierungsproblem ist die Minimierung einer konvexen Funktion über einer konvexen Menge. Empirische Risikominimierung (ERM) ist ein konvexes Optimierungsproblem, wenn der Hypothesenraum durch eine konvexe Menge von Modellparameter parametrisiert ist und die Verlustfunktion in diesen Modellparameter konvex ist. Jeder lokale Minimierer einer konvexen Zielfunktion ist notwendigerweise auch ein globaler Minimierer, sodass ein Verfahren, das ein lokales Minimum erreicht, das Optimierungsproblem gelöst hat.

Definition

Das Lernen eines linearen Modell durch Gradientenabstieg bewegt die Modellparameter schrittweise, wobei jede Iteration die Zielfunktion in einer Umgebung der aktuellen Modellparameter verringert. Viele ML- Methoden lernen Modellparameter durch eine solche iterative Optimierungsverfahren, wobei Gradientenabstieg und seine Varianten am weitesten verbreitet sind.

Diese Optimierungsverfahren konvergieren garantiert zu einem globalen Minimum, wenn die Zielfunktion eine differenzierbar konvexe Funktion der Modellparameter ist, die ihr Minimum annimmt und eine Lipschitz-Stetigkeit stetig Gradient besitzt (siehe glatt), und die Schrittweite passend gewählt wird (Boyd and Vandenberghe, 2004, Sekt. 9.3; Bubeck, 2015, Thm. 3.3; Nesterov, 2004, Thm. 2.1.14). Für eine konvexe Funktion ist jedes lokale Minimum notwendigerweise ein globales Minimum: Jede Wahl von Modellparameter, bei der keine kleine Änderung die Zielfunktion weiter verringert, erreicht bereits den kleinsten Wert, den die Zielfunktion irgendwo annimmt. Der mittlere quadratischer Verlust der linearen Regression ist eine konvexe Funktion der Modellparameter, ebenso wie die Zielfunktion der Stützvektormaschine (SVM). Empirische Risikominimierung (ERM) über einem Hypothesenraum, der durch eine konvexe Menge von Modellparameter parametrisiert ist, mit einer Verlustfunktion, die in diesen konvex ist, ist ein konvexes Optimierungsproblem: die Minimierung einer konvexen Funktion über einer konvexen Menge.

Konvexität ist sowohl für Mengen als auch für Funktionen definiert. Für Mengen ist die Definition geometrisch. Eine Teilmenge $\mathcal{C} \subseteq \mathbb{R}^d$ des euklidischen Raum $\mathbb{R}^d$ heißt konvex, wenn sie das Streckensegment zwischen beliebigen zwei ihrer Punkte enthält (Boyd and Vandenberghe, 2004, Sekt. 2.1.4). Für $\mathbf{w}, \mathbf{w}' \in \mathcal{C}$ gilt also

$$\beta\mathbf{w} + (1-\beta)\mathbf{w}' \in \mathcal{C} \quad \text{für alle } \beta \in [0,1].$$

Analog ist eine Funktion $f : \mathbb{R}^d \to \mathbb{R}$ konvex, wenn ihr Epigraph $\left\{ \left(\mathbf{w}^\top, t\right)^\top \in \mathbb{R}^{d+1} \,:\, t \geq f(\mathbf{w}) \right\}$ eine konvexe Menge ist (Boyd and Vandenberghe, 2004, Sekt. 3.1.7), wobei der Epigraph die Menge der Punkte auf oder über dem Graph einer Funktion von f ist. Beispiele für eine konvexe Menge und eine konvexe Funktion sind in Abb. 1 dargestellt.



(a) (b)

Abbildung 1: (a) Konvexe Menge $\mathcal{C} \subseteq \mathbb{R}^d$: Das Streckensegment zwischen beliebigen zwei Punkten von $\mathcal{C}$ bleibt in $\mathcal{C}$. (b) Konvexe Funktion $f : \mathbb{R}^d \to \mathbb{R}$: der schattierte Epigraph ist eine konvexe Menge, und das Streckensegment, das beliebige zwei Punkte auf dem Graph einer Funktion von f verbindet, liegt oberhalb des Graph einer Funktion

Zwei Familien konvexer Mengen treten in der maschinellen Lernen immer wieder auf. Die erste ist der Halbraum $\{\mathbf{w} \in \mathbb{R}^d : \mathbf{a}^\top \mathbf{w} \leq b\}$. Er wird durch einen von Null verschiedenen Normalenvektor $\mathbf{a} \in \mathbb{R}^d \backslash \{\mathbf{0}\}$ bestimmt, der die Richtung der begrenzenden Hyperebene festlegt, und einen Offset $b \in \mathbb{R}$, der ihre Position festlegt. Wenn zwei Punkte die Ungleichung erfüllen, tut dies auch jeder Punkt des Segments zwischen ihnen (Boyd and Vandenberghe, 2004, Sekt. 2.2.1). Die zweite ist die konvexe Hülle endlich vieler Vektoren $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}$. Sie ist definiert als die Menge aller ihrer konvexen Kombinationen,

$$\operatorname{conv}\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)}\right\} := \left\{ \sum_{r=1}^{m} \beta_r \mathbf{x}^{(r)} \;:\; \beta_r \geq 0 \text{ für } r = 1, \ldots, m, \; \sum_{r=1}^{m} \beta_r = 1 \right\}.$$

Sie ist die kleinste konvexe Menge, die diese Vektoren enthält (Boyd and Vandenberghe, 2004, Sekt. 2.1.4).

Jede abgeschlossene konvexe Menge ist der Schnitt aller Halbräume, die sie enthalten (Boyd and Vandenberghe, 2004, Sekt. 2.3.1; Rockafellar, 1970,

Thm. 11.5)

$$\mathcal{C} = \bigcap \big\{ H \subseteq \mathbb{R}^d \ : \ H \text{ ist ein Halbraum, } \mathcal{C} \subseteq H \big\}. \tag{1}$$

Im Allgemeinen ergibt der Schnitt einer Teilmenge der Halbräume, die eine konvexe Menge enthalten, eine konvexe Menge, die die ursprüngliche enthält, d. h. eine äußere Approximation (siehe Abb. 2).

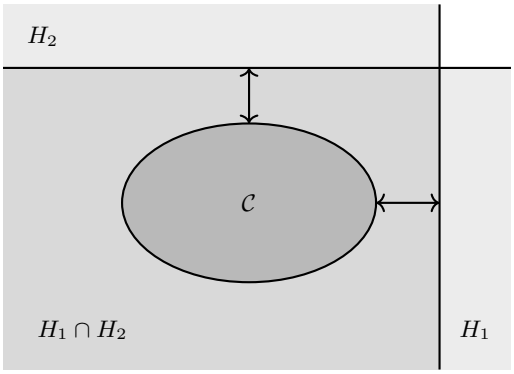

Abbildung 2: Eine konvexe Menge $\mathcal{C}$ und zwei Halbräume H_1, H_2, die sie enthalten, in drei Grautönen: am hellsten, wo nur ein Halbraum die Ebene abdeckt, mittelgrau für ihren Schnitt $H_1 \cap H_2$, am dunkelsten für $\mathcal{C}$ selbst. Keiner der Halbräume ist eng: Seine begrenzende Hyperebene berührt $\mathcal{C}$ nicht, sodass die durch die Doppelpfeile markierten Lücken bleiben. Der Schnitt ist daher eine strikt größere konvexe Menge als $\mathcal{C}$ – eine äußere Approximation

Der Halbraum-Schnitt (1) lässt sich einschränken: Es müssen nur die Halbräume betrachtet werden, die von einer Stützhyperebene begrenzt werden. Jeder Rand- Punkt einer nichtleeren konvexen Menge trägt eine solche Hyperebene (Boyd and Vandenberghe, 2004, Sekt. 2.5.2), und jeder Punkt außerhalb der Menge wird bereits durch einen der Halbräume ausgeschlossen, die diese Stützhyperebenen begrenzen: Für einen Punkt $\mathbf{x} \notin \mathcal{C}$ und $\mathcal{C}$ abgeschlossen gibt es einen eindeutig nächsten Punkt $\mathbf{p} \in \mathcal{C}$ zu $\mathbf{x}$ (Boyd and Vandenberghe, 2004, Sekt. 8.1). Dieser nächste Punkt ist ein Rand-Punkt von $\mathcal{C}$ und damit einer der Punkte, die eine Stützhyperebene tragen: Ein Punkt im Inneren könnte ein wenig in Richtung $\mathbf{x}$ verschoben werden und läge immer noch in $\mathcal{C}$, wäre also nicht der nächste. Die Hyperebene durch $\mathbf{p}$ mit Normalenvektor $\mathbf{x} - \mathbf{p}$, also orthogonal zu $\mathbf{x} - \mathbf{p}$, stützt $\mathcal{C}$ an $\mathbf{p}$, und der von ihr begrenzte Halbraum enthält $\mathcal{C}$, aber nicht $\mathbf{x}$ (Boyd and Vandenberghe, 2004, Sekt. 2.5.1 und 2.5.2; siehe Abb. 3). Abb. 4 zeigt acht Stützhyperebenen einer Ellipse.

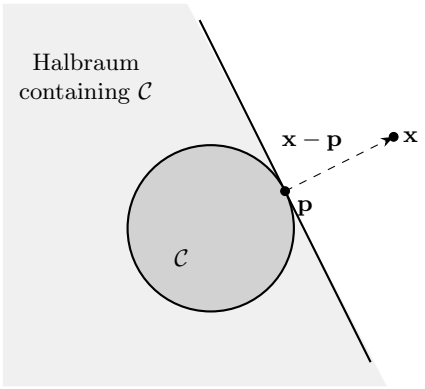

Abbildung 3: Warum die Stützhyperebenen bereits ausreichen, um eine abgeschlossene konvexe Menge $\mathcal{C}$ wiederherzustellen, hier eine Kreisscheibe (dunklere Schattierung). Für einen Punkt $\mathbf{x}$ außerhalb von $\mathcal{C}$ sei $\mathbf{p}$ der Punkt in $\mathcal{C}$, der $\mathbf{x}$ am nächsten liegt. Die Hyperebene durch $\mathbf{p}$, die orthogonal zu $\mathbf{x} - \mathbf{p}$ ist (durchgezogene Linie), berührt $\mathcal{C}$ in $\mathbf{p}$, ist also eine Stützhyperebene, und der von ihr begrenzte Halbraum (hellere Schattierung) enthält $\mathcal{C}$, aber nicht $\mathbf{x}$. Jeder Punkt außerhalb von $\mathcal{C}$ wird von einem solchen Halbraum ausgeschlossen, sodass der Schnitt dieser Halbräume $\mathcal{C}$ wiederherstellt.

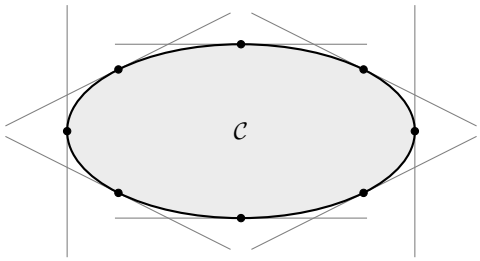

Abbildung 4: Eine konvexe Menge $\mathcal{C}$ (schattierte Ellipse) und acht Hyperebenen, die sie berühren (graue Linien). Jede von ihnen begrenzt einen Halbraum, der $\mathcal{C}$ enthält, und der Schnitt aller Halbräume, die $\mathcal{C}$ enthalten, ist $\mathcal{C}$ selbst.

Für den Epigraph einer differenzierbar konvexen Funktion f wird die Stützhyperebene durch den Gradient bestimmt. Am Rand-Punkt $\left(\mathbf{w}^\top, f(\mathbf{w})\right)^\top$ ist ihr Normalenvektor $\left(\nabla f(\mathbf{w})^\top, -1\right)^\top$, denn

$$\left(\nabla f(\mathbf{w})^\top, -1\right) \left[\begin{pmatrix} \mathbf{w}' \\ t \end{pmatrix} - \begin{pmatrix} \mathbf{w} \\ f(\mathbf{w}) \end{pmatrix} \right] \leq 0 \quad \text{für alle } \left(\mathbf{w}'^\top, t\right)^\top \in \operatorname{epi} f.$$

Dies ist die Bedingung erster Ordnung $f(\mathbf{w}') \geq f(\mathbf{w}) + \nabla f(\mathbf{w})^\top (\mathbf{w}' - \mathbf{w})$, ausgedrückt als Aussage über den Epigraph (Boyd and Vandenberghe, 2004,

Sekt. 3.1.7). An einer Stelle, an der der Gradient verschwindet, ergibt die Bedingung $f(\mathbf{w}') \geq f(\mathbf{w})$ für jedes $\mathbf{w}'$, was der obigen Aussage entspricht, dass ein lokales Minimum einer konvexen Funktion ein globales ist.

Die äußere Approximation aus Abb. 2 wird auch von approximativen Inferenz-Verfahren in Wahrscheinlichkeitsmodelle verwendet, die über einem Graph definiert sind. Die Menge der erreichbaren Randverteilungen ist eine konvexe Menge, die durch eine große Zahl von linearen Ungleichungen beschrieben wird. Message-Passing-Verfahren optimieren über eine äußere Approximation davon, die aus einer Teilmenge dieser Ungleichungen gewonnen wird (Wainwright and Jordan, 2008, Sekt. 3.4.1; Sontag and Jaakkola, 2007).

Von den beiden oben eingeführten Familien konvexer Mengen ist die konvexe Hülle diejenige, die die lineare Separierbarkeit von Datenpunkte mit binären Label-Werten charakterisiert. Betrachtet werden m Datenpunkte mit Merkmalsvektoren $\mathbf{x}^{(r)} \in \mathbb{R}^d$ und Labels $y^{(r)} \in \{-1, +1\}$, für $r = 1, \ldots, m$. Diese Datenpunkte sind linear separierbar, wenn ein linearer Klassifikator jedes Label korrekt vorhersagt, d. h. es existieren $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ und $b \in \mathbb{R}$ mit $y^{(r)}\big(\mathbf{w}^\top \mathbf{x}^{(r)} + b\big) > 0$ für $r = 1, \ldots, m$. Lineare Separierbarkeit liegt genau dann vor, wenn sich die konvexe Hülle der mit $+1$ gelabelten Merkmalsvektoren und die konvexe Hülle der mit -1 gelabelten nicht schneiden (Boyd and Vandenberghe, 2004, Sekt. 2.5.1 und 8.6.1). Separierbarkeit ist also eine Eigenschaft zweier aus den Datenpunkte gebildeter konvexer Mengen.
Konvexität tritt auch in einer weit verbreiteten Familie von Wahrscheinlichkeitsmodelle auf. Eine Exponentialfamilie besteht aus den Wahrscheinlichkeitsverteilungen, deren Wahrscheinlichkeitsdichte die Form

$$p(\mathbf{x}; \mathbf{w}) = h(\mathbf{x}) \exp\big(\mathbf{w}^\top \mathbf{t}(\mathbf{x}) - A(\mathbf{w})\big).$$

durchläuft. Diese Wahrscheinlichkeitsdichte ist parametrisiert durch den Vektor $\mathbf{w}$ und die suffizienten Statistiken $\mathbf{t}(\cdot)$. Die Log-Partitions-Funktion A ist eindeutig dadurch bestimmt, dass $p(\mathbf{x}; \mathbf{w})$ auf eins integriert. Ihre Hesse-Matrix ist die Kovarianzmatrix der suffizienten Statistiken $\mathbf{t}(\mathbf{x})$ unter $p(\mathbf{x}; \mathbf{w})$,

$$\nabla^2 A(\mathbf{w}) = \mathbb{E}\big\{\mathbf{t}(\mathbf{x})\mathbf{t}(\mathbf{x})^\top\big\} - \mathbb{E}\big\{\mathbf{t}(\mathbf{x})\big\}\mathbb{E}\big\{\mathbf{t}(\mathbf{x})\big\}^\top,$$

die positiv semidefinit ist. Eine zweimal-differenzierbar Funktion, deren Hesse-Matrix positiv semidefinit ist, ist konvex (Boyd and Vandenberghe, 2004, Sekt. 3.1.4), also ist A eine konvexe Funktion auf ihrer Definitionsbereich (Wainwright and Jordan, 2008, Prop. 3.1). Die Konvexität von A impliziert, dass der negative Log-Likelihood-Funktion eines Trainingsdatensatz, bis auf eine additive Konstante, die nicht von $\mathbf{w}$ abhängt,

$$-\sum_{r=1}^{m} \mathbf{w}^\top \mathbf{t}\big(\mathbf{x}^{(r)}\big) + mA(\mathbf{w}),$$

eine konvexe Funktion in $\mathbf{w}$ ist. Für eine Exponentialfamilie ist die maximale Likelihood-Schätzung somit ein konvexes Optimierungsproblem.

Siehe auch: Epigraph, Halbraum, Stützhyperebene, konvexe Optimierung, Hesse-Matrix.

Querverweise

Literatur

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

2. Bubeck (2015). Convex Optimization: Algorithms and Complexity. Found. Trends Mach. Learn., 8(3–4), pp. 231–357.

3. Wainwright and Jordan (2008). Graphical Models, Exponential Families, and Variational Inference. Found. Trends Mach. Learn., 1(1–2), pp. 1–305.

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

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

6. Sontag and Jaakkola (2007). New Outer Bounds on the Marginal Polytope. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 1393–1400.