quadratischer Verlust

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 quadratische Verlust einer Hypothese auf einem Datenpunkt mit einem numerischen Label ist das Quadrat der Differenz zwischen dem Label und der Vorhersage, die die Hypothese für die Merkmale des Datenpunkts liefert. Sein Mittelwert über eine Trainingsdatensatz ist der Trainingsfehler, auch mittlerer quadratischer Fehler (MSE) genannt, und die Minimierung dieses Mittelwerts über ein lineares Modell ist die kleinste Quadrate, deren Lösung die Normalgleichungen in geschlossener Form angeben. Bis auf eine streng monoton wachsende affine Transformation ist der quadratische Verlust die negative Log-Likelihood-Funktion des Labels unter gaußschem Rauschen um die Vorhersage, sodass die kleinste Quadrate eine maximale Likelihood-Schätzung ist. Die Hypothese mit dem kleinsten Risiko unter dem quadratischen Verlust ist der bedingter Erwartungswert des Labels gegeben die Merkmale, und das Risiko jeder anderen Hypothese übersteigt dieses um die erwartete quadratische Abweichung von diesem bedingter Erwartungswert. Der quadratische Verlust ist konvex und differenzierbar in der Vorhersage mit einer zum Fehler proportionalen Steigung, sodass ein einzelner Ausreißer die gelernte Hypothese proportional zu seinem Fehler verschiebt, während der absoluter Verlust und der Huber-Verlust diese Verschiebung begrenzen.

Definition

Ein Wetterdienst sagt die Lufttemperatur um 15:00 Uhr (UTC) in Krems an der Donau (Österreich) aus dem Messwert um 06:00 Uhr (UTC) desselben Tages voraus. Jeder Tag ist ein Datenpunkt mit dem Merkmal x, dem Messwert um 06:00 Uhr, und dem Label $y \in \mathbb{R}$, dem Messwert um 15:00 Uhr. Eine Hypothese h liefert die Vorhersage $\widehat{y} = h(x)$, und die um 15:00 Uhr gemessene Temperatur gibt an, um wie viele Grad diese Vorhersage danebenlag. Für einen Datenpunkt mit dem Merkmalsvektor $\mathbf{x} \in \mathcal{X}$ und dem numerischen Label $y \in \mathbb{R}$ werde eine Hypothese $h : \mathcal{X} \to \mathbb{R}$ betrachtet. Der quadratische Verlust, den sie erleidet, ist das Quadrat der Differenz zwischen dem Label und ihrer Vorhersage (Hastie et al., 2009, Abschn. 2.4),

$$L\left((\mathbf{x}, y), h\right) := \big(y - h(\mathbf{x})\big)^2. \tag{1}$$

Am 9. Dezember 2024 betrug der Messwert um 06:00 Uhr in Krems 2.6 Grad und der Messwert um 15:00 Uhr 3.8 Grad. Die weiter unten gelernte lineare Hypothese $\widehat{h}(x) = 6.08 + 1.00x$ sagt für diesen Tag 8.68 Grad voraus, sodass ihr quadratischer Verlust (1) darauf $(3.8 - 8.68)^2 \approx 23.8$ Quadratgrad beträgt. Der quadratische Verlust wird in der quadrierten Einheit des Labels gemessen, verschwindet genau dann, wenn die Vorhersage korrekt ist, und ist für eine um denselben Betrag zu hohe Vorhersage genauso groß wie für eine zu niedrige.

Als Funktion der Vorhersage $\widehat{y}$ ist der quadratische Verlust (1) eine Parabel mit ihrem Minimum bei $\widehat{y} = y$, wie Abb. 1 zeigt. Er ist konvex und differenzierbar mit der Ableitung $-2(y - \widehat{y})$: Seine Steigung wächst proportional zum Fehler. Für ein lineares Modell $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ ist der quadratische Verlust ebenfalls eine konvexe und differenzierbare Funktion der Modellparameter $\mathbf{w}$, was das Gradientenabstieg anwendbar macht. Der absoluter Verlust $|y - \widehat{y}|$ ist ebenfalls konvex, aber nichtglatt bei $\widehat{y} = y$, und seine Steigung beträgt ± 1, unabhängig von der Größe des Fehlers. Der Huber-Verlust mit Schwellenwert $\delta > 0$ verläuft quadratisch für einen kleinen Fehler und linear für einen großen (Hastie et al., 2009, Gl. (10.23)). Ausgedrückt durch den Fehler $e := y - h(\mathbf{x})$ lautet er

$$\begin{cases} e^2 & \text{für } |e| \le \delta, \\ 2\delta|e| - \delta^2 & \text{sonst,} \end{cases} \tag{2}$$

sodass er für einen Fehler bis zu δ mit dem quadratischen Verlust übereinstimmt und seine Steigung darüber hinaus bei 2δ gedeckelt ist (Huber, 1992).

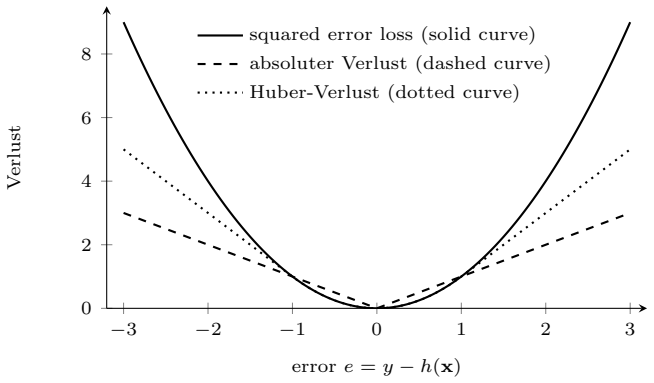


Abbildung 1: Der quadratische Verlust (1) (durchgezogene Kurve), der absoluter Verlust (gestrichelte Kurve) und der Huber-Verlust (2) mit Schwellenwert $\delta = 1$ (gepunktete Kurve) als Funktionen des Fehlers $e = y - h(\mathbf{x})$. Die drei stimmen bei einem Fehler von ± 1 überein; darüber hinaus wird nur der quadratische Verlust immer steiler

Der Mittelwert des quadratischen Verlusts über eine Trainingsdatensatz

$\mathcal{D}^{(\text{train})} = \left\{ \left(\mathbf{x}^{(r)}, y^{(r)}\right) \right\}_{r=1}^{m}$,

$$\frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - h(\mathbf{x}^{(r)})\right)^2, \tag{3}$$

ist der Trainingsfehler von h, auch ihr mittlerer quadratischer Fehler (MSE) genannt, und die empirische Risikominimierung (ERM) wählt aus einem Modell die Hypothese mit dem kleinsten Trainingsfehler aus. Mit einem lineares Modell ist die ERM unter dem quadratischen Verlust eine lineare Regression mittels der kleinste Quadrate (Legendre, 1805). Der Trainingsfehler (3) lautet dann $(1/m)\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$, mit der Merkmalsmatrix $\mathbf{X} \in \mathbb{R}^{m \times d}$, deren Zeilen die Merkmalsvektoren sind, und dem Labelvektor $\mathbf{y} \in \mathbb{R}^m$. Sein Gradient verschwindet genau an den Lösungen der Normalgleichungen $\mathbf{X}^\top \mathbf{X}\mathbf{w} = \mathbf{X}^\top \mathbf{y}$, und wenn $\mathbf{X}$ vollen Spalten-Rang hat, ist die eindeutige Lösung (Bishop, 2006, Abschn. 3.1.1)

$$\widehat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{4}$$

Die geschlossene Form (4) benötigt keine Iteration, sodass es keine Aktualisierungsregel gibt, deren Fixpunkt die Lösung wäre. Wenn $\mathbf{X}^\top \mathbf{X}$ zu groß zum Invertieren ist, wird stattdessen das Gradientenabstieg auf dem Trainingsfehler (3) verwendet, und dessen Fixpunkte sind wiederum die Lösungen der Normalgleichungen (siehe lineare Regression).

Als Trainingsdatensatz werde jeder 18. Tag des Jahres 2024 in Krems gewählt, für den beide Messwerte vorliegen (21 Tage), wobei das konstante Merkmal 1 an den Messwert um 06:00 Uhr angehängt wird, sodass $\widehat{\mathbf{w}}$ einen Achsenabschnitt und eine Steigung enthält. Die geschlossene Form (4) liefert die Gerade $\widehat{h}(x) = 6.08 + 1.00x$ aus Abb. 2(a) mit einem Trainingsfehler von 10.22 Quadratgrad beziehungsweise 3.20 Grad nach Ziehen der Wurzel. Die beste konstante Vorhersage unter dem quadratischen Verlust ist der Stichprobenmittel der Labels, 14.75 Grad, und ihr Trainingsfehler ist deren Stichprobenvarianz, 79.47: Der mittlere quadratische Fehler einer Konstanten c ist diese Varianz plus $(c - 14.75)^2$. Unter dem absoluter Verlust ist die beste Konstante stattdessen der Median der Labels, 14.50 Grad.

Der quadratische Verlust ist der Verlust hinter der maximale Likelihood-Schätzung, wenn das Label als Vorhersage plus gaußsches Rauschen modelliert wird. Sei $y = h(\mathbf{x}) + \varepsilon$ mit einer Gaußsche Zufallsvariable ε mit Mittelwert null und Varianz σ^2, sodass y gegeben $\mathbf{x}$ die Wahrscheinlichkeitsdichte $\mathcal{N}\left(h(\mathbf{x}), \sigma^2\right)$ besitzt. Der negative Logarithmus dieser Wahrscheinlichkeitsdichte am beobachteten Label ist (Bishop, 2006, Abschn. 3.1.1, Gln. (3.11)–(3.12))

$$-\ln \mathcal{N}\left(h(\mathbf{x}), \sigma^2\right)(y) = \frac{1}{2\sigma^2}\left(y - h(\mathbf{x})\right)^2 + \frac{1}{2}\ln\left(2\pi\sigma^2\right), \tag{5}$$

eine streng monoton wachsende affine Funktion des quadratischen Verlusts (1). Für Datenpunkte, die Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) sind, ist die negative Log-Likelihood-Funktion der Trainingsdatensatz die Summe der Terme (5) über ihre Datenpunkte, was dem

$m/(2\sigma^2)$-Fachen des Trainingsfehlers (3) plus einer Konstanten entspricht. Der maximale Likelihood-Schätzwert für $\mathbf{w}$ ist daher die Lösung der kleinste Quadrate (4), unabhängig vom Wert von σ^2, und der maximale Likelihood-Schätzwert für σ^2 ist der Trainingsfehler selbst, 10.22 für die 21 Tage. Der logistische Verlustfunktion ist die negative Log-Likelihood-Funktion für ein binäres Label, und Laplace-Rauschen anstelle von gaußschem Rauschen führt auf den absoluter Verlust (Murphy, 2012, Abschn. 7.4).
Das Risiko einer Hypothese ist ihr erwarteter Verlust unter einem Wahrscheinlichkeitsmodell mit einer gemeinsamen Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{x},y)}$ des Merkmalsvektors und des Labels. Unter dem quadratischen Verlust wird das Risiko durch den bedingter Erwartungswert $h^*(\mathbf{x}) = \mathbb{E}\{y \mid \mathbf{x}\}$ minimiert, den Bayes-Schätzer für diesen Verlust (Bishop, 2006, Abschn. 1.5.5, Gl. (1.89); Hastie et al., 2009, Abschn. 2.4, Gl. (2.13)). Das Risiko jeder Hypothese h zerfällt in zwei Teile (Bishop, 2006, Gl. (1.90)),

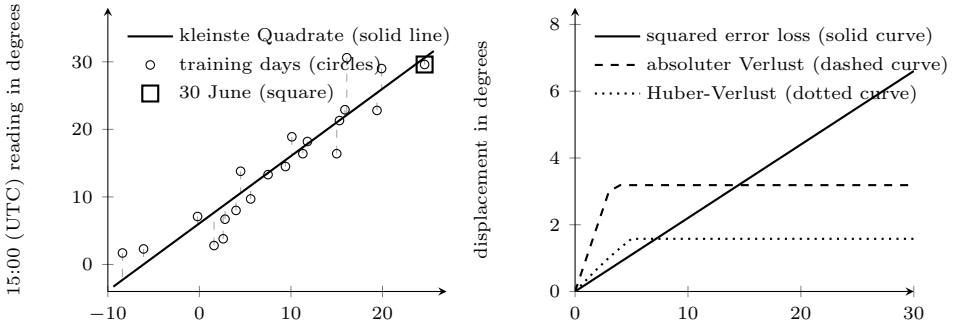
$$\mathbb{E}\big\{\big(y - h(\mathbf{x})\big)^2\big\} = \mathbb{E}\big\{\big(h^*(\mathbf{x}) - h(\mathbf{x})\big)^2\big\} + \mathbb{E}\big\{\big(y - h^*(\mathbf{x})\big)^2\big\}, \tag{6}$$

die erwartete quadrierte Differenz zwischen h und dem bedingter Erwartungswert sowie die erwartete bedingte Varianz des Labels, die keine Hypothese reduzieren kann. Für eine aus einem zufälligen Trainingsdatensatz gelernte Hypothese zerfällt der erste Teil der Zerlegung (6) weiter in die quadrierte Verzerrung und die Varianz der gelernten Vorhersage, die Zerlegung hinter dem Bias-Varianz-Dilemma (Hastie et al., 2009, Abschn. 7.3, Gl. (7.9)). Unter dem absoluter Verlust ist der Bayes-Schätzer der bedingte Median (Hastie et al., 2009, Gl. (2.18)). Die beiden konstanten Vorhersagen oben sind die Trainingsdatensatz-Versionen dieser beiden Tatsachen, mit dem Mittelwert über die 21 Tage anstelle des Erwartungswerts.

Die wachsende Steigung der Parabel hat ihren Preis. Ein Datenpunkt, dessen Label weit von der Vorhersage entfernt liegt, geht proportional zu seinem Fehler in den Gradienten des Trainingsfehlers (3) ein, sodass ein Ausreißer die gelernte Hypothese unbeschränkt zu sich hinzieht (Hastie et al., 2009, Abschn. 10.6). Bei der kleinste Quadrate ist dieser Zug explizit. Addiert man Δ zum Label des r-ten Trainingstages, ändert sich die Lösung (4) um $(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{(r)}\Delta$. Die Vorhersage für diesen Tag ändert sich um $\ell_r\Delta$, wobei $\ell_r = (\mathbf{x}^{(r)})^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^{(r)}$ die Hebelwirkung (Leverage) des Tages ist.

Unter den 21 Tagen hat derjenige mit dem wärmsten 06:00-Uhr-Messwert, der 30. Juni 2024 mit 24.6 Grad um 06:00 Uhr und 29.6 Grad um 15:00 Uhr, die Hebelwirkung 0.220. Jedes zu seinem Label addierte Grad verschiebt die Gerade der kleinste Quadrate bei diesem Messwert um 0.220 Grad, was bei $\Delta = 30$ einem Wert von 6.60 Grad entspricht, und die Steigung der Geraden steigt von 1.00 auf 1.32 (Abb. 2(b), durchgezogene Kurve). Unter dem absoluter Verlust, dessen ERM-Lösung die Kleinste-Absolute-Abweichungs-Regression ist, verschiebt sich die Gerade um höchstens 3.19 Grad und überhaupt nicht mehr, sobald Δ den Wert 4 überschreitet (gestrichelte Kurve). Die Steigung des absoluter Verlusts beträgt ± 1; sobald der verfälschte Tag also oberhalb der Geraden liegt, hängt sein Term in der Optimalitätsbedingung nicht mehr von der Größe

seines Fehlers ab. Der Huber-Verlust mit $\delta = 3$ verschiebt seine Gerade um 1.58 Grad und jenseits von $\Delta = 5$ überhaupt nicht mehr (gepunktete Kurve), da seine Steigung gedeckelt ist. Dieser beschränkte Zug macht die Robustheit gegenüber Ausreißern für einen Verlust aus. Der absoluter Verlust ist nichtglatt, und der Huber-Verlust führt den Schwellenwert δ als zusätzlichen Hyperparameter ein; der quadratische Verlust hat keinen dieser Nachteile, aber auch keine solche Schranke.



(a) the 21 training days and the kleinste Quadrate line (b) displacement of the learned line at that day

Abbildung 2: (a) Jeder 18. Tag des Jahres 2024 in Krems, die 21 Trainingstage (Kreise) und die mittels der kleinste Quadrate gelernte Gerade (durchgezogene Linie): Der quadratische Verlust (1) eines Tages ist die quadrierte Länge seines gestrichelten Segments, und der Trainingsfehler (3) mittelt diese über die 21 Tage. Das Quadrat markiert den 30. Juni, dessen Label in (b) verfälscht wird. (b) Die Verschiebung der gelernten Geraden beim 06:00-Uhr-Messwert dieses Tages, wenn Δ Grad zu seinem Label addiert werden: Die Gerade der kleinste Quadrate (durchgezogene Kurve) verschiebt sich unbeschränkt um 0.220Δ, während die unter dem absoluter Verlust (gestrichelte Kurve) und unter dem Huber-Verlust mit Schwellenwert $\delta = 3$ (gepunktete Kurve) gelernten Geraden sich nicht weiter verschieben, sobald der verfälschte Tag weit genug oberhalb von ihnen liegt. Daten erzeugt mit `pythondemos/sqerrloss.py`

Wenn das Label ein Vektor ist, ist der quadratische Verlust die quadrierte euklidische Norm der Differenz zwischen Label und Vorhersage. Der Clusteringfehler von k-Means, der einen Datenpunkt mit dem ihn repräsentierenden Clusterzentrum vergleicht, sowie der Rekonstruktionsfehler der Hauptkomponentenanalyse (PCA) und eines Autoencoders haben diese Form, wobei der Merkmalsvektor selbst an die Stelle des Labels tritt.

Synonyme: quadratische Fehlerfunktion.

Siehe auch: Verlust, Verlustfunktion, absoluter Verlust, Huber-Verlust, logistische Verlustfunktion, 0/1-Verlust, Trainingsfehler, MSE, ERM, kleinste Quadrate, lineare Regression, Normalgleichungen, Gradientenabstieg, maximale Likelihood, Gaußsche Zufallsvariable, Risiko, Bayes-Schätzer, bedingter

Erwartungswert, Bias-Varianz-Dilemma, Ausreißer, Robustheit, Kleinste-Absolute-Abweichungs-Regression, Huber-Regression, Clusteringfehler.

Querverweise

Literatur

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

2. Legendre (1805). Nouvelles méthodes pour la détermination des orbites des comètes. Firmin Didot, Paris.

3. Murphy (2012). Machine Learning: A Probabilistic Perspective. MIT Press, Cambridge, MA, USA.

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.

5. Huber (1992). Robust estimation of a location parameter. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 492–518. Springer-Verlag.