0/1-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 0/1-Verlust eines Klassifikators auf einem Datenpunkt ist 1, wenn das vorhergesagte Label vom wahren Label abweicht, und 0, wenn beide übereinstimmen; er ist somit die Indikatorfunktion einer Fehlklassifikation. Sein Durchschnitt über einen Datensatz ist der Anteil der fehlklassifizierten Datenpunkte, was eins minus der Genauigkeit entspricht, und sein Erwartungswert ist die Wahrscheinlichkeit einer Fehlklassifikation, deren kleinster Wert über alle Klassifikatoren das Bayes-Risiko ist. Bei der binäre Klassifikation hängt er vom Datenpunkt nur über den Margin ab, und er nutzt die reellwertige Hypothese nur über ihr Vorzeichen. Als Funktion der Modellparameter ist er stückweise konstant mit einem Gradienten gleich null überall dort, wo der Gradient existiert, und das Finden des linearer Klassifikators mit den wenigsten Fehlklassifikationen auf einem Trainingsdatensatz ist NP-schwer, sodass er nicht als Zielfunktion beim Training verwendet wird. Die empirische Risikominimierung (ERM) verwendet stattdessen einen konvexen Ersatzverlust, den Hinge-Verlust oder den durch $\ln 2$ geteilten logistische Verlustfunktion, die jeweils eine obere Schranke für den 0/1-Verlust darstellen und deren Minimierer über alle Funktionen das Vorzeichen des Klassifikators mit dem Bayes-Risiko haben. Der 0/1-Verlust bleibt der Verlust, anhand dessen ein Klassifikator auf einem Testdatensatz bewertet wird.

Definition

Eine Frostwarnung für Krems an der Donau wird am Abend ausgegeben und ist am nächsten Morgen entweder richtig oder falsch: Es gab Frost oder nicht. Eine falsche Warnung als 1 und eine richtige als 0 zu zählen, ist der 0/1-Verlust. Im Allgemeinen betrachte man einen Datenpunkt mit Merkmalen $\mathbf{x} \in \mathcal{X}$ und einem Label y aus einem endlichen Labelraum $\mathcal{Y}$ sowie einen Klassifikator $h : \mathcal{X} \to \mathcal{Y}$, der das Label vorhersagt. Der 0/1-Verlust, den h auf dem Datenpunkt $(\mathbf{x}, y)$ erleidet, ist (Goodfellow et al., 2016, Abschn. 5.1.2; Shalev-Shwartz and Ben-David, 2014, Abschn. 3.2.2)

$$L^{(0/1)}\left((\mathbf{x}, y), h\right) := \begin{cases} 0 & \text{falls } h(\mathbf{x}) = y, \\ 1 & \text{falls } h(\mathbf{x}) \neq y. \end{cases} \tag{1}$$

Er ist die Indikatorfunktion einer Fehlklassifikation und nimmt dieselben zwei Werte an, unabhängig davon, ob der Labelraum zwei oder viele Elemente enthält.

Der Durchschnitt des 0/1-Verlustes (1) über einen Datensatz $\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \ldots, (\mathbf{x}^{(m)}, y^{(m)})\}$ ist der Anteil der fehlklassifizierten Datenpunkte,

$$\frac{1}{m} \sum_{r=1}^{m} L^{(0/1)} \left((\mathbf{x}^{(r)}, y^{(r)}), h \right) = 1 - \text{acc}(h \mid \mathcal{D}), \tag{2}$$

was eins minus der Genauigkeit von h auf $\mathcal{D}$ ist. Auf einem Trainingsdatensatz ist der Durchschnitt (2) der Trainingsfehler von h unter dem 0/1-Verlust, und auf einem Testdatensatz ist er die für eine Methode zur Klassifikation angegebene Fehlerrate. Für Datenpunkte, die Realisierungen von unabhängig und identisch verteilt (u.i.v.) Zufallsvariablen (ZV) mit einer gemeinsamen Wahrscheinlichkeitsverteilung sind, ist der Erwartungswert des 0/1-Verlustes die Wahrscheinlichkeit, dass h einen Datenpunkt fehlklassifiziert, was das Risiko von h unter dem 0/1-Verlust ist (Shalev-Shwartz and Ben-David, 2014, Abschn. 3.2.2). Das Risiko ist am kleinsten für den Klassifikator, der jedem $\mathbf{x}$ den Label-Wert mit der größten bedingten Wahrscheinlichkeit gegeben $\mathbf{x}$ zuweist, und dieses kleinste Risiko ist das Bayes-Risiko (Bishop, 2006, Abschn. 1.5.1).

Bei der binäre Klassifikation mit $\mathcal{Y} = \{-1, +1\}$ wird ein Klassifikator üblicherweise aus einer reellwertigen Hypothese $h : \mathcal{X} \to \mathbb{R}$ gebildet, die $\widehat{y} = \text{sign}(h(\mathbf{x}))$ vorhersagt. Der 0/1-Verlust (1) dieses Klassifikators hängt vom Datenpunkt nur über den Margin $y \cdot h(\mathbf{x})$ ab,

$$L^{(0/1)}\left((\mathbf{x}, y), h\right) = \begin{cases} 0 & \text{falls } y \cdot h(\mathbf{x}) > 0, \\ 1 & \text{falls } y \cdot h(\mathbf{x}) \leq 0. \end{cases} \tag{3}$$

Ein positiver Margin ist eine korrekte Vorhersage, und ein Datenpunkt auf der Entscheidungsgrenze $h(\mathbf{x}) = 0$ gilt als fehlklassifiziert. Die Margin-Form (3) verwendet h nur über ihr Vorzeichen, sodass die Multiplikation von h mit einer beliebigen positiven Zahl den 0/1-Verlust jedes Datenpunkts unverändert lässt.

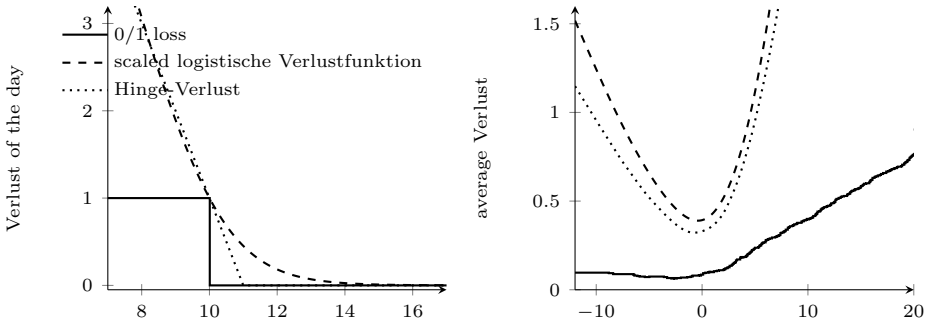
Der 0/1-Verlust ist der natürliche Verlust für die Klassifikation, und die empirische Risikominimierung (ERM) damit würde den Trainingsfehler (2) als die über einen Hypothesenraum $\mathcal{H}$ zu minimierende Zielfunktion verwenden. Zwei Hindernisse stehen dem im Weg. Erstens ist der durchschnittliche 0/1-Verlust als Funktion der Modellparameter von h stückweise konstant: Er ändert sich nur, wenn die Entscheidungsgrenze einen Datenpunkt überquert, und zwischen zwei solchen Überquerungen ist sein Gradient null. Der Gradientenabstieg und seine Varianten erhalten daher keine Verbesserungsrichtung aus ihm, und der Durchschnitt ist weder konvex noch glatt (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Zweitens ist das Problem selbst für Methoden schwer, die keinen Gradienten verwenden. Wenn irgendein linearer Klassifikator jeden Datenpunkt des Trainingsdatensatzs korrekt klassifiziert, findet ein lineares Programm einen solchen (Shalev-Shwartz and Ben-David, 2014, Abschn. 9.1). Ohne diese Garantie ist das Finden des linearer Klassifikators mit den wenigsten Fehlklassifikationen auf einem gegebenen Trainingsdatensatz NP-schwer, und ebenso das Finden

eines solchen, dessen Anzahl korrekter Vorhersagen innerhalb eines festen Faktors des größtmöglichen Wertes liegt (Ben-David et al., 2003; Shalev-Shwartz and Ben-David, 2014, Abschn. 8.2.2).

Abb. 1 zeigt beide Hindernisse anhand der eingangs erwähnten Frostwarnung. Jeder der $m = 363$ Tage des Jahres 2024 in Krems, dessen Folgetag erfasst ist, ist ein Datenpunkt, dessen einziges Merkmal x die Lufttemperatur um 06:00 Uhr (UTC) ist und dessen Label $y = +1$ lautet, wenn der Messwert um 06:00 Uhr des folgenden Tages unter 0 °C liegt (Frost), und andernfalls $y = -1$. Auf 35 der 363 Tage folgt Frost. Die Warnung vergleicht den Messwert des Tages um 06:00 Uhr mit einem Schwellenwert τ: Die Hypothese $h(x) = \tau - x$ sagt Frost vorher, wenn sie positiv ist, und der einzige Modellparameter τ wählt einen Klassifikator aus dem Hypothesenraum aller Schwellenwerte aus. Der 0/1-Verlust eines Tages ist eine Stufe in τ, von 1 auf 0 beim Messwert des Tages, und der Durchschnitt über das Jahr ist eine Treppenfunktion mit 177 Stufen, eine bei jedem unterschiedlichen Messwert. Ihr kleinster Wert, $23/363 = 0.063$, wird von jedem Schwellenwert strikt zwischen -2.9 und -2.7 sowie zwischen -2.6 und -2.4 °C angenommen: Die Menge der Minimierer ist kein einzelner Punkt, und die Steigung der Treppenfunktion ist bei jedem Schwellenwert, der kein Sprung ist, gleich null. Eine Warnung, die niemals Frost vorhersagt, hat einen durchschnittlichen 0/1-Verlust von $35/363 = 0.096$, und eine, die immer Frost vorhersagt, hat 0.904.

Praktische Methoden zur Klassifikation ersetzen den 0/1-Verlust in der ERM durch eine Ersatz-Verlustfunktion, die konvex in $h(\mathbf{x})$ und eine obere Schranke für den 0/1-Verlust ist (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Der Hinge-Verlust $\max\{0, 1-y\cdot h(\mathbf{x})\}$, der auf die Stützvektormaschine (SVM) führt, liegt für jeden Margin auf oder über dem 0/1-Verlust (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Der logistische Verlustfunktion $\ln(1+\exp(-y\cdot h(\mathbf{x})))$, der auf die logistische Regression führt, ist beim Margin 0 gleich $\ln 2 \approx 0.69$ und erst nach Division durch $\ln 2$ eine obere Schranke. Einen Ersatzverlust so zu skalieren, dass er beim Margin 0 durch 1 verläuft, ist die übliche Konvention beim Vergleich von Verlustfunktionen (Bishop, 2006, Abb. 7.5; Hastie et al., 2009, Abschn. 10.6). Da der Ersatzverlust eine obere Schranke ist, beträgt das Risiko jeder Hypothese unter dem 0/1-Verlust höchstens ihr Risiko unter dem Ersatzverlust, sodass eine Hypothese mit einem kleinen Ersatz-Risiko selten fehlklassifiziert (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Beide Ersatzverluste sind zudem für die Klassifikation kalibriert: Für eine konvexe Funktion des Margins, die bei 0 differenzierbar ist und dort eine negative Ableitung hat, weist jede Folge von Hypothesen, deren Ersatz-Risiko gegen seinen kleinstmöglichen Wert strebt, ein 0/1-Risiko auf, das gegen das Bayes-Risiko strebt (Bartlett et al., 2006, Sätze 1 und 6). Über alle Funktionen von $\mathbf{x}$ wird der Hinge-Verlust durch das Vorzeichen von $\mathbb{P}(y = +1 \mid \mathbf{x}) - 1/2$ minimiert und der logistische Verlustfunktion durch den Logarithmus der Chancen von $y = +1$, und das Vorzeichen jedes der beiden Minimierer ist der Klassifikator mit dem Bayes-Risiko (Hastie et al., 2009, Tab. 12.1). Im Gegensatz zum 0/1-Verlust ändert sich ein Ersatzverlust, wenn h mit einer positiven Zahl multipliziert wird; er hängt also von der Skalierung von h ab und nicht nur von ihrem Vorzeichen.

Für die Frostwarnung sind der durch $\ln 2$ geteilte durchschnittliche logistische Verlustfunktion und der durchschnittliche Hinge-Verlust konvexe Funktionen des Schwellenwerts, die überall über der Treppenfunktion liegen (Abb. 1). Jede hat einen einzigen Minimierer, $\tau = -0.40\,°\mathrm{C}$ für den logistische Verlustfunktion und $-0.70\,°\mathrm{C}$ für den Hinge-Verlust, und bei beiden beträgt der durchschnittliche 0/1-Verlust $29/363 = 0.080$, sechs Tage mehr als der kleinste Wert der Treppenfunktion: Das Minimieren eines Ersatzverlusts ist nicht das Minimieren des 0/1-Verlustes. Beim Minimierer des logistische Verlustfunktiones beträgt der durch $\ln 2$ geteilte durchschnittliche logistische Verlustfunktion 0.390 für h, 0.285 für $h/2$ und 0.685 für $2h$, während der durchschnittliche 0/1-Verlust für alle drei gleich ist.



(a) one day: 2 November, $x = 10.0\,°\mathrm{C}$ (b) average over the 363 days

Abbildung 1: Die Frostwarnung $h(x) = \tau - x$ in Krems als Funktion ihres Schwellenwerts τ. (a) Der 0/1-Verlust (durchgezogene Kurve), der logistische Verlustfunktion geteilt durch $\ln 2$ (gestrichelte Kurve) und der Hinge-Verlust (gepunktete Kurve) eines Tages, des 2. November 2024, dessen Messwert um 06:00 Uhr (UTC) $x = 10.0\,°\mathrm{C}$ betrug und auf den Frost folgte. Der 0/1-Verlust springt bei $\tau = 10.0$ von 1 auf 0, wo beide Ersatzverluste gleich 1 sind, und die Ersatzverluste liegen bei jedem Schwellenwert auf oder über dem 0/1-Verlust. (b) Die Mittelwerte derselben drei Verluste über die 363 Tage des Jahres 2024. Der mittlere 0/1-Verlust ist eine Treppenfunktion mit 177 Stufen und seinem kleinsten Wert von 0.063 knapp unter $-2.5\,°\mathrm{C}$; die beiden gemittelten Ersatzverluste sind konvexe Kurven mit jeweils einem Minimierer bei -0.40 und $-0.70\,°\mathrm{C}$, etwa zwei Grad über den Minimierern der Treppenfunktion. Daten erzeugt mit `pythondemos/zerooneloss.py`

Ein kleiner durchschnittlicher 0/1-Verlust auf dem Trainingsdatensatz ist nicht das Ziel des Trainings, sondern ein Mittel zu einem kleinen Risiko, und beide können voneinander abweichen. Abb. 2 zeigt ein Trainingsdatensatz, auf dem ein linearer Klassifikator h' einen Datenpunkt fehlklassifiziert, während ein Klassifikator h'', dessen Entscheidungsgrenze sich um diesen Datenpunkt windet, keinen fehlklassifiziert. Eine kleine Verschiebung eines Merkmalsvektors nahe dieser Entscheidungsgrenze kehrt die Vorhersage von h'' um, sodass sein

Trainingsfehler von null nicht auf eine gute Generalisierung hindeutet; dies ist Überanpassung. Das Einschränken des Hypothesenraums auf lineare Klassifikatoren ist eine Form der Regularisierung, jenes Weges, der den Hypothesenraum verkleinert.

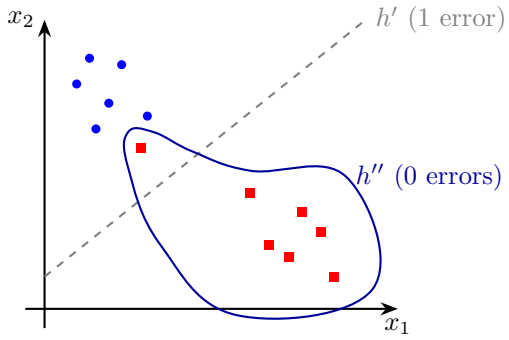


Abbildung 2: Zwei Klassifikatoren auf einer Trainingsdatensatz aus Datenpunkte mit zwei Merkmale, dargestellt als blaue Kreise und rote Quadrate entsprechend ihrem Labelwert. Der linearer Klassifikator h' (gestrichelte Linie) klassifiziert einen Datenpunkt falsch, das rote Quadrat zwischen den blauen Kreisen, sodass sein mittlerer 0/1-Verlust auf der Trainingsdatensatz 1/13 beträgt. Der Klassifikator h'' (durchgezogene Kurve) krümmt seine Entscheidungsgrenze um diesen Datenpunkt herum und erreicht einen mittleren 0/1-Verlust von null, doch bereits eine kleine Verschiebung eines Merkmalsvektors nahe dieser Entscheidungsgrenze kehrt seine Vorhersage um

Synonyme: Fehlklassifikationsverlust, Null-Eins-Verlust.

Siehe auch: Verlust, Verlustfunktion, Klassifikation, binäre Klassifikation, Klassifikator, Genauigkeit, Trainingsfehler, Risiko, Bayes-Risiko, Margin, Hinge-Verlust, logistische Verlustfunktion, ERM, SVM, logistische Regression, Konfusionsmatrix, unausgewogene Daten.

Querverweise

Literatur

1. Bartlett et al. (2006). Convexity, Classification, and Risk Bounds. Journal of the American Statistical Association, 101(473), pp. 138–156.

2. Ben-David et al. (2003). On the Difficulty of Approximately Maximizing Agreements. Journal of Computer and System Sciences, 66(3), pp. 496–514.

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

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

5. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

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