logistische Verlustfunktion

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

Die logistische Verlustfunktion ist eine konvexe und glatte Ersatzfunktion für den 0/1-Verlust bei der binäre Klassifikation. Wie der Hinge-Verlust hängt sie von einem Datenpunkt nur über dessen Margin ab, das Produkt aus seinem Label und dem Wert der Hypothese. Im Gegensatz zum Hinge-Verlust ist sie überall differenzierbar, wodurch sie sich für Gradientenbasierte Optimierungsverfahren wie das Gradientenabstieg eignet. Die Anwendung der empirische Risikominimierung (ERM) mit einem lineares Modell und dem logistischen Verlust ergibt die logistische Regression. Der logistische Verlust ist der negative Logarithmus der bedingte Wahrscheinlichkeitsverteilung des Labels gegeben den Merkmalsvektor, sodass seine Minimierung einer maximale Likelihood-Schätzung der Modellparameter entspricht. Ein Wahrscheinlichkeitsmodell, in dem die Merkmale gegeben das Label gemeinsam Gaußsche Zufallsvariablen sind, induziert genau diese bedingte Wahrscheinlichkeitsverteilung, und es ist nicht das einzige, das dies tut.

Definition

Ein Spamfilter ist ein Verfahren der binäre Klassifikation, das eine Hypothese erlernt, welche die Merkmale einer E-Mail entweder auf *Spam* oder *kein Spam* abbildet. Die Güte eines solchen Spamfilters lässt sich messen, indem der Anteil der falsch klassifizierten E-Mails gezählt wird. Dieser Anteil ist der durchschnittliche 0/1-Verlust. Allerdings verläuft dieser durchschnittliche Verlust als Funktion der Modellparameter fast überall flach, sodass er dem Gradientenabstieg keine Abstiegsrichtung vorgibt. Seine Form entspricht der in Abb. 1 dargestellten Stufe: flach auf beiden Seiten der Entscheidungsgrenze und nirgendwo mit einer Steigung versehen, der man folgen könnte. Praktische Verfahren zur Klassifikation verwenden daher eine konvexe und glatte Ersatzfunktion für den 0/1-Verlust, die überall eine von null verschiedene Steigung aufweist. Eine solche Ersatzfunktion ist der logistische Verlust, der im Folgenden definiert wird.

Betrachtet sei ein Problem der binäre Klassifikation mit dem Merkmalsraum $\mathcal{X}$ und einer reellwertigen Hypothese $h : \mathcal{X} \to \mathbb{R}$. Die beiden Klassen werden im Folgenden durch das Label $y \in \{-1, +1\}$ kodiert. Der logistische Verlust, den h auf einem Datenpunkt $(\mathbf{x}, y)$ erleidet, beträgt (Shalev-Shwartz and Ben-David,

2014, Abschn. 9.3, S. 127)

$$L\left((\mathbf{x}, y), h\right) := \ln\left(1 + \exp(-y \cdot h(\mathbf{x}))\right). \tag{1}$$

Wie der Hinge-Verlust hängt auch der logistische Verlust (1) vom Datenpunkt nur über den Margin $y \cdot h(\mathbf{x})$ ab. Eine Normierung des logistischen Verlustes durch $\ln 2$ macht ihn zu einer konvexen oberen Schranke für den 0/1-Verlust, mit Gleichheit beim Margin 0 (siehe Abb. 1). Erst die Umskalierung macht ihn zu einer Schranke: Beim Margin 0 beträgt der logistische Verlust selbst $\ln 2 \approx 0.69$ und liegt damit unter dem dortigen Wert 1 des 0/1-Verlustes. Eine Ersatzfunktion so zu skalieren, dass sie beim Margin 0 durch 1 verläuft, ist die übliche Konvention, wenn solche Verlustfunktionen miteinander verglichen werden (Hastie et al., 2009, Abschn. 10.6).

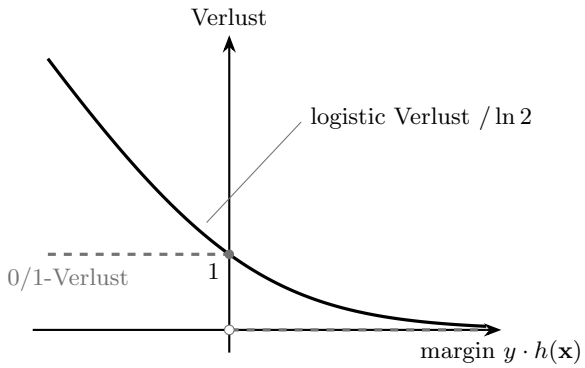

Abbildung 1: Der logistische Verlust $L\left((\mathbf{x}, y), h\right)$ geteilt durch $\ln 2$ (durchgezogene Kurve) und der 0/1-Verlust (gestrichelte Kurve) als Funktionen des Margins $y \cdot h(\mathbf{x})$. Der umskalierte logistische Verlust liegt für jeden Margin auf oder über dem 0/1-Verlust, mit Gleichheit beim Margin 0; die Umskalierung ist erforderlich, da $\ln\left(1 + \exp(0)\right) = \ln 2 < 1$ gilt

Im Gegensatz zum Hinge-Verlust ist der logistische Verlust überall differenzierbar. Er eignet sich daher gut für Gradientenbasierte Optimierungsverfahren wie das Gradientenabstieg. Die empirische Risikominimierung (ERM) mit einem lineares Modell und dem logistischen Verlust führt zur logistische Regression.

Der logistische Verlust ergibt sich auf natürliche Weise aus einem einfachen Wahrscheinlichkeitsmodell, in dem die Datenpunkte unabhängig und identisch verteilt (u.i.v.)e Zufallsvariablen (ZV) mit einer gemeinsamen Verbund-Wahrscheinlichkeitsverteilung $\mathbb{P}^{(\mathbf{x},y)}$ von Merkmalsvektor und Label sind. Bedingt auf das Label y ist der Merkmalsvektor $\mathbf{x}$ ein Gaußsche Zufallsvariable mit dem Mittelwert-Vektor $\boldsymbol{\mu}^{(-1)}$ für $y = -1$ und $\boldsymbol{\mu}^{(+1)}$ für $y = +1$. Die bedingte Kovarianzmatrix ist für beide gleich und fest auf $\mathbf{I}$ gesetzt.

Unter diesem Wahrscheinlichkeitsmodell ist die A-posteriori-Wahrscheinlichkeit von $y = +1$ durch die Sigmoidfunktion $\sigma(h(\mathbf{x}))$ mit $\sigma(z) = 1/(1 + e^{-z})$ ge-

geben. Diese Identität gilt für die affine Hypothese $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$ mit den Modellparametern (Bishop, 2006, Abschn. 4.2.1, Gl. (4.65)–(4.67))

$$\mathbf{w} = \boldsymbol{\mu}^{(+1)} - \boldsymbol{\mu}^{(-1)}, \quad b = -\tfrac{1}{2}\left(\left\|\boldsymbol{\mu}^{(+1)}\right\|_2^2 - \left\|\boldsymbol{\mu}^{(-1)}\right\|_2^2\right) + \ln \frac{\mathbb{P}(y=+1)}{\mathbb{P}(y=-1)}.$$

Es ist die gemeinsame Kovarianzmatrix, die h affin macht: Die in $\mathbf{x}$ quadratischen Terme sind unter beiden Labelsn gleich und kürzen sich im Quotienten heraus, sodass nur das Skalarprodukt mit $\mathbf{w}$ übrig bleibt. Wenn beide Labels gleich wahrscheinlich sind, verschwindet der Logarithmus, und b besteht allein aus dem Mittelpunktsterm. Die Wahrscheinlichkeit von $y = -1$ ist dann $\sigma(-h(\mathbf{x}))$, da die Sigmoidfunktion die Gleichung $\sigma(-z) = 1 - \sigma(z)$ erfüllt (Bishop, 2006, Gl. 4.60). Der logistische Verlust ist somit der negative Logarithmus der bedingten Wahrscheinlichkeit des Labels gegeben den Merkmalsvektor, und die Minimierung des durchschnittlichen logistischen Verlustes entspricht einer maximale Likelihood-Schätzung des GewichtsVektors $\mathbf{w}$ und des Versatzes b von h unter dieser bedingte Wahrscheinlichkeitsverteilung (Bishop, 2006, Abschn. 4.3.2; Murphy, 2012, Abschn. 8.3). Das Gauß'sche Wahrscheinlichkeitsmodell ist nicht das einzige, das diese bedingte Wahrscheinlichkeitsverteilung induziert. Besitzt der Merkmalsvektor gegeben das Label eine Dichte der Form $h(\mathbf{x}) \exp\left((\mathbf{w}^{(y)})^\top \mathbf{t}(\mathbf{x}) - A(\mathbf{w}^{(y)})\right)$, gehört er also zu einer Exponentialfamilie, deren natürlicher Parameter $\mathbf{w}^{(y)}$ vom Label abhängt, während das Basis-Maß h und die suffiziente Statistik $\mathbf{t}(\mathbf{x})$ dies nicht tun, so ist der Logarithmus des A-posteriori-Chancenverhältnisses von $y = +1$ affin in $\mathbf{t}(\mathbf{x})$, und die A-posteriori-Wahrscheinlichkeit ist wiederum die Sigmoidfunktion einer affinen Funktion (Bishop, 2006, Abschn. 4.2.4). Ihr GewichtsVektor ist die Differenz der beiden natürlichen Parameter, $\mathbf{w} = \mathbf{w}^{(+1)} - \mathbf{w}^{(-1)}$, was für das obige Gauß'sche Modell, bei dem $\mathbf{t}(\mathbf{x}) = \mathbf{x}$ gilt und der natürliche Parameter der Mittelwert-Vektor ist, genau der dort angegebenen Identität $\mathbf{w} = \boldsymbol{\mu}^{(+1)} - \boldsymbol{\mu}^{(-1)}$ entspricht. Bedingt unabhängige binäre Merkmale, das Modell hinter naiver Bayes, bilden einen weiteren Fall (Bishop, 2006, Abschn. 4.2.3).

Ein Problem der binäre Klassifikation, dessen Labelraum ein anderes Paar $\{a, b\}$ ist, wird durch eine Kodierung $e : \{a, b\} \to \{-1, +1\}$ mit $e(a) = -1$ und $e(b) = +1$ abgedeckt. Der logistische Verlust (1) wird dann am kodierten Label ausgewertet,

$$L\left((\mathbf{x}, y), h\right) = \ln\left(1 + \exp(-e(y) \cdot h(\mathbf{x}))\right).$$

Die Labelräume $\{0, 1\}$ und $\{\textit{no spam}, \textit{spam}\}$ werden beide auf diese Weise behandelt. Betrachtet sei der Labelraum $\{0, 1\}$ mit $e(0) = -1$ und $e(1) = +1$. Der Verlust lässt sich dann ausschreiben als

$$-y \ln \sigma\big(h(\mathbf{x})\big) - (1 - y) \ln\big(1 - \sigma(h(\mathbf{x}))\big),$$

was dem binären Kreuzentropie-Verlust (engl. *binary cross-entropy loss*, BCE) entspricht: der Kreuzentropie zwischen dem Label und der vorhergesagten Wahrscheinlichkeit $\sigma(h(\mathbf{x}))$, wobei das Label 1 den zweiten Summanden zu null macht und das Label 0 den ersten. Summiert über ein Trainingsdatensatz ergibt

dies die Kreuzentropie-Fehlerfunktion der logistische Regression (Bishop, 2006, Gl. 4.90).

In dieser Schreibweise setzt der Verlust nicht mehr voraus, dass das Label einen von zwei Werten annimmt. Sei $\tilde{y}$ ein Soft Label, also eine Zahl zwischen 0 und 1, welche die Wahrscheinlichkeit angibt, dass der Datenpunkt zu der durch 1 kodierten Klasse gehört. Setzt man $\tilde{y}$ anstelle von y ein, so erhält man den Kreuzentropie-Verlust

$$-\tilde{y} \ln \sigma\big(h(\mathbf{x})\big) - (1-\tilde{y}) \ln \big(1-\sigma(h(\mathbf{x}))\big),$$

der sich für $\tilde{y} \in \{0,1\}$ auf den vorherigen Ausdruck reduziert und sich mittels der Softmax-Funktion auf mehr als zwei Klassen übertragen lässt.

Der Wert des logistischen Verlustes ist schwerer zu interpretieren als der 0/1-Verlust, da er vom Betrag des Margins und nicht nur von dessen Vorzeichen abhängt. Jede im Folgenden verglichene Hypothese ist ein linearer Klassifikator $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b$. Betrachtet seien die drei Datenpunkte aus Abb. 2, zwei davon mit dem Label $+1$ bei $(1,1)$ und $(2,3)$ sowie einer mit dem Label -1 bei $(3,1)$, und dazu der linearer Klassifikator $\widehat{h}^{(1)}(\mathbf{x}) = -0.1x_1 + 0.05x_2 + 0.15$. Seine Margins betragen bei jedem der drei Punkte 0.1, sodass er alle drei korrekt klassifiziert und einen durchschnittlichen logistischen Verlust von etwa 0.64 erleidet. Multipliziert man seine Modellparameter mit 50, so erhält man $50\widehat{h}^{(1)}$, wodurch die Entscheidungsgrenze exakt an derselben Stelle bleibt und sich somit keine Vorhersage ändert, während der durchschnittliche logistische Verlust jedoch unter 0.01 fällt. Zwei Hypothesen, deren logistischer Verlust weit auseinanderliegt, können also exakt dieselben Entscheidungen treffen.
Der logistische Verlust und der 0/1-Verlust können zwei Hypothesen auch in entgegengesetzter Reihenfolge bewerten. Ein zweiter linearer Klassifikator $\widehat{h}^{(2)}$ besitzt eine zur ersten parallele, aber verschobene Entscheidungsgrenze, wodurch der Datenpunkt bei $(3,1)$ auf der falschen Seite zu liegen kommt. Seine Margins betragen 5, 5 und -0.5, was einer Genauigkeit von 2/3 gegenüber 1 für $\widehat{h}^{(1)}$ entspricht, und dennoch erzielt er den kleineren durchschnittlichen logistischen Verlust von etwa 0.33 gegenüber 0.64. Seine beiden großen Margins verursachen fast keine Kosten, sodass selbst der einzelne negative Margin, der etwa 1 kostet, in der Summe weniger ergibt als die drei kleinen Margins von $\widehat{h}^{(1)}$. Eine Auswahl zwischen den beiden allein anhand des logistischen Verlustes würde also das ungenauere Modell bevorzugen. Nur eine einzige Interpretation des Zahlenwerts ist verlässlich, nämlich die obige obere Schranke: Ein kleiner logistischer Verlust erzwingt einen kleinen 0/1-Verlust.

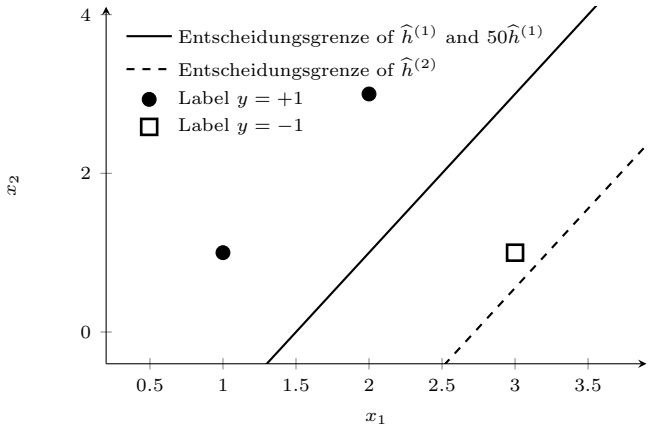

Abbildung 2: Die drei Datenpunkte des Beispiels, wobei die zwei mit dem Label $y = +1$ als ausgefüllte Kreise und der eine mit dem Label $y = -1$ als offenes Quadrat dargestellt sind, sowie die Entscheidungsgrenzen der auf ihnen verglichenen Hypothesen. Skaliert man $\widehat{h}^{(1)}$ mit 50, werden ihre Modellparameter mit 50 multipliziert, sodass die Entscheidungsgrenze genau dort bleibt, wo sie war (durchgezogene Linie): Jede Vorhersage bleibt unverändert, während der durchschnittliche Wert der logistischen Verlust sinkt. Die Entscheidungsgrenze von $\widehat{h}^{(2)}$ (gestrichelte Linie) verläuft parallel, ist aber verschoben, wodurch der Datenpunkt bei $(3, 1)$ auf der falschen Seite liegt. Daten erzeugt mit `pythondemos/logloss.py`

Synonyme: logistischer Verlust, Log-Loss, binärer Kreuzentropieverlust.

Siehe auch: 0/1-Verlust, Hinge-Verlust, Margin, Genauigkeit, Kreuzentropie, Wahrscheinlichkeitsmodell, Sigmoidfunktion, Klassifikator, logistische Regression, Gradientenabstieg, konvex, binäre Klassifikation.

Querverweise

Literatur

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

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

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