binäre Klassifikation

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 binäre Klassifikation bezeichnet Anwendungen des maschinellen Lernens mit einem binären Labelraum. Das Ziel der binären Klassifikation besteht darin, jeden Datenpunkt einer von zwei Klassen zuzuordnen, wobei jede Klasse durch einen der beiden Label-Werte repräsentiert wird. Streng genommen ist ein binärer Klassifikator eine Hypothese-Abbildung vom Merkmalsraum in einen binären Labelraum. Es ist jedoch üblich, einen solchen Klassifikator zu konstruieren, indem der Wert einer reellwertigen Hypothese-Abbildung mit einem Schwellenwert verglichen wird. Lineare Klassifikatoren ergeben sich aus dem Vorzeichen einer linearen Hypothese-Abbildung. Ein binärer Klassifikator unterteilt den Merkmalsraum in zwei Entscheidungsregionen, die durch eine Entscheidungsgrenze getrennt sind, welche bei einem linearen Klassifikator eine Hyperebene bildet. Die natürliche Verlustfunktion ist der 0/1-Verlust, dessen Durchschnitt über einen Datensatz gleich 1 minus der Genauigkeit ist. Allerdings ist diese Verlustfunktion für Methoden des Trainings schwer zu optimieren, und ihr Durchschnitt auf unausgewogene Daten wird vom Label-Wert der Mehrheitsklasse dominiert, sodass er keinen Aufschluss darüber gibt, welche der beiden Fehlerarten ein Klassifikator begeht. Praktische Verfahren verwenden daher häufig andere Verlustfunktionen für das Training und die Validierung eines Modells.

Definition

Ein Frostwarnsystem sagt voraus, ob die Temperatur in der kommenden Nacht unter 0 °C fällt. Dies ist ein Beispiel für die binäre Klassifikation, das Problem des maschinellen Lernens, aus den Merkmale eines Datenpunkts vorherzusagen, welchen von zwei Werten sein Label, eine Zahl y, annimmt. Bei dem Frostwarnsystem repräsentiert ein Datenpunkt einen Tag, der durch die Lufttemperatur um 06:00 (UTC) und um 15:00 (UTC) dieses Tages als seine beiden Merkmale charakterisiert ist. Der Label-Wert gibt an, ob der Messwert um 06:00 (UTC) des folgenden Morgens unter 0 °C fällt oder nicht.

Im Allgemeinen ist die binäre Klassifikation eine Klassifikation mit einem binären Labelraum $\mathcal{Y}$. Jeder Datenpunkt trägt ein Label, und dieses Label nimmt einen der beiden Werte an. Die beiden Werte werden üblicherweise durch Zahlen wie $\mathcal{Y} = \{-1, +1\}$ oder $\mathcal{Y} = \{0, 1\}$ dargestellt (oder kodiert). Die Wahl

der Kodierung ist eine Konvention, doch mathematische Ausdrücke setzen meist eine davon voraus, am häufigsten $\{-1, +1\}$, und müssen für die andere angepasst werden (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3; Hastie et al., 2009, Abschn. 4.2).

Streng genommen ist ein binärer Klassifikator eine Hypothese-Abbildung, die Werte in einem binären Labelraum $\mathcal{Y}$ annimmt. Er wird üblicherweise aus einer reellwertigen Hypothese-Abbildung $h : \mathcal{X} \to \mathbb{R}$ gebildet, die dem Merkmalsvektor $\mathbf{x}$ eines Datenpunkts die Zahl $h(\mathbf{x})$ zuordnet. Hierbei ist $\mathcal{X}$ der Merkmalsraum, die Menge aller möglichen Merkmalsvektoren. Der binäre Klassifikator vergleicht diese Zahl mit einem Schwellenwert. Für die Label-Kodierung $\mathcal{Y} = \{-1, +1\}$ und den Schwellenwert 0 ist der vorhergesagte Label-Wert des Datenpunkts das Vorzeichen $\widehat{y} = \text{sign}(h(\mathbf{x}))$ (Shalev-Shwartz and Ben-David, 2014, Abschn. 9.1).

Ein binärer Klassifikator unterteilt den Merkmalsraum $\mathcal{X}$ in zwei Entscheidungsregionen, eine pro Label-Wert: die Merkmalsvektoren mit $\widehat{y} = +1$ und diejenigen mit $\widehat{y} = -1$. Die beiden Entscheidungsregionen werden durch die Entscheidungsgrenze getrennt, die Menge der Merkmalsvektoren $\mathbf{x}$ mit $h(\mathbf{x}) = 0$. Das Produkt $y \cdot h(\mathbf{x})$ aus dem Label y eines Datenpunkts und dem Wert von h an seinem Merkmalsvektor ist der Margin von h auf diesem Datenpunkt (Shalev-Shwartz and Ben-David, 2014, Abschn. 15.1).

Für einen linearen Klassifikator lautet die reellwertige Hypothese $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_0$. Ihre Modellparameter sind der Vektor $\mathbf{w}$, der einen Eintrag für jedes Merkmal enthält, und die Zahl w_0. Die Entscheidungsgrenze ist dann die Hyperebene $\mathbf{w}^\top \mathbf{x} + w_0 = 0$, also eine Gerade bei zwei Merkmale (Bishop, 2006, Abschn. 4.1.1). Die Vorhersage wird durch die Seite dieser Hyperebene bestimmt, auf der ihr Merkmalsvektor $\mathbf{x}$ liegt.

Ein linearer Klassifikator verursacht Klassifikationsfehler für einen Datensatz, wenn dieser Datenpunkte mit demselben Label enthält, die auf beiden Seiten der Entscheidungsgrenze liegen. Idealerweise sollten sich Datenpunkte derselben Klasse (mit identischen Label-Werten) alle auf derselben Seite der Entscheidungsgrenze befinden. Darüber hinaus sollten diese Datenpunkte im Sinne der Robustheit auch weit von der Entscheidungsgrenze entfernt sein. Der Abstand eines Datenpunkts mit dem Merkmalsvektor $\mathbf{x}$ von der Entscheidungsgrenze eines linearen Klassifikators beträgt $|h(\mathbf{x})|/\|\mathbf{w}\|$. Werden die Modellparameter so skaliert, dass $\|\mathbf{w}\| = 1$ gilt, entspricht der Margin von h auf einem korrekt klassifizierten Datenpunkt diesem Abstand (Shalev-Shwartz and Ben-David, 2014, Abschn. 15.1).

Die natürliche Verlustfunktion für die binäre Klassifikation ist der 0/1-Verlust: Er weist einer fehlerhaften Klassifikation ($\widehat{y} \neq y$) den Verlust 1 und einer korrekten Klassifikation ($\widehat{y} = y$) den Verlust 0 zu. Sein Durchschnitt über einen Datensatz ist daher der Anteil der falsch klassifizierten Datenpunkte, und 1 minus diesem Anteil ist die Genauigkeit. Einer der 8 Datenpunkte in Abb. 1 ist falsch klassifiziert, sodass der durchschnittliche 0/1-Verlust dort 1/8 und die Genauigkeit 7/8 beträgt.

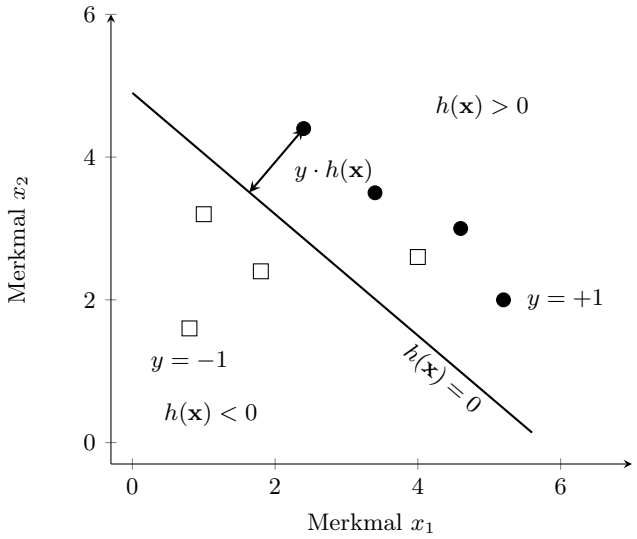
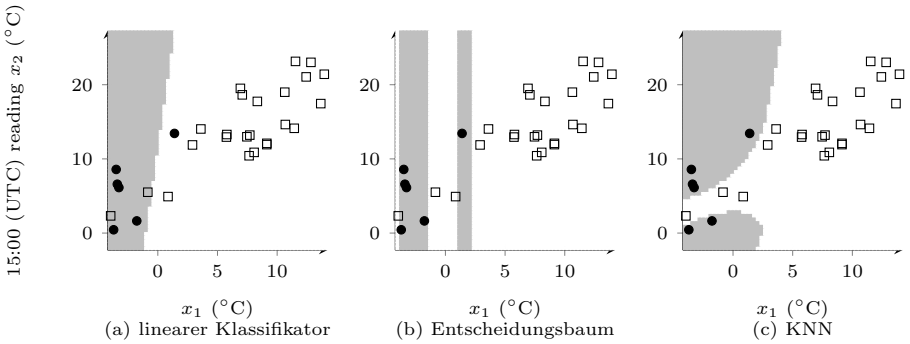


Abbildung 1: Binäre Klassifikation von Datenpunkten mit zwei Merkmale unter Verwendung eines linearen Klassifikators. Die Entscheidungsgrenze $h(\mathbf{x}) = 0$ ist eine Hyperebene, hier eine Gerade, und sie teilt den Merkmalsraum in die beiden Entscheidungsregionen der Label-Werte $y = +1$ (ausgefüllte Kreise, wo $h(\mathbf{x}) > 0$) und $y = -1$ (offene Quadrate, wo $h(\mathbf{x}) \leq 0$). Die Modellparameter sind so skaliert, dass $\|\mathbf{w}\| = 1$ gilt, und der Doppelpfeil zeigt den Margin $y \cdot h(\mathbf{x})$ eines Datenpunkts, der dann seinem Abstand von der Geraden entspricht. Einer der 8 Datenpunkte ist falsch klassifiziert, das offene Quadrat oberhalb der Geraden: Er verursacht einen 0/1-Verlust von 1, sodass der durchschnittliche 0/1-Verlust 1/8 und die Genauigkeit 7/8 beträgt

Die Form der Entscheidungsregionen eines binären Klassifikators $\widehat{h}$ hängt von dem Modell ab, das zum Lernen des Klassifikators $\widehat{h}$ verwendet wird. Abb. 2 zeigt drei Klassifikatoren, die aus demselben Trainingsdatensatz von 30 Datenpunkten gelernt wurden. Der lineare Klassifikator teilt den Merkmalsraum in zwei Halbebenen und klassifiziert drei der 30 Tage falsch. Der Entscheidungsbaum leitet einen Datenpunkt durch eine Folge von Verzweigungen, von denen jede ein Merkmal mit einem Schwellenwert vergleicht und den Datenpunkt an einen von zwei Ästen weitergibt. Eine Verzweigung teilt somit eine Region entlang einer Linie parallel zu einer Merkmal-Achse, und die Entscheidungsregionen des Entscheidungsbaums sind Vereinigungen achsenparalleler Rechtecke (Hastie et al., 2009, Abschn. 9.2.1). Ein Blatt ist ein Knoten ohne weitere Verzweigung darunter, und es enthält den Label-Wert, der für die Datenpunkte, die es erreichen, vorhergesagt wird. Wird der Entscheidungsbaum so weit aufgebaut, bis jedes Blatt nur noch von Datenpunkten mit einem einzigen Label-Wert erreicht wird, sagt er den Label-Wert jedes Datenpunkts im Trainingsdatensatz korrekt voraus. Das künstliche neuronale Netz verwendet tanh-Aktivierungsfunktionen,

die glatt sind, sodass seine Entscheidungsgrenze eine glatte nichtlineare Kurve ist. Wie der Entscheidungsbaum liefert auch das künstliche neuronale Netz korrekte Vorhersagen für alle Datenpunkte im Trainingsdatensatz.



(a) linearer Klassifikator (b) Entscheidungsbaum (c) KNN

Abbildung 2: Entscheidungsregionen dreier binärer Klassifikatoren, die aus denselben 30 Datenpunkten gelernt wurden. Jeder Datenpunkt repräsentiert die Wetterbedingungen rund um einen bestimmten Tag: Seine Merkmale sind die Lufttemperatur um 06:00 (UTC) und um 15:00 (UTC) dieses Tages in Grad Celsius. Ausgefüllte Kreise sind Tage, auf die Frost folgt ($y = +1$), offene Quadrate die übrigen Tage ($y = -1$). Der schattierte Bereich ist die Entscheidungsregion, in der der Klassifikator +1 vorhersagt. (a) Ein linearer Klassifikator, dessen Regionen zwei Halbebenen sind. (b) Ein Entscheidungsbaum, der so weit aufgebaut wurde, bis die Datenpunkte, die jedes Blatt erreichen, einen einzigen Label-Wert tragen, und dessen Regionen Vereinigungen achsenparalleler Rechtecke sind. (c) Ein KNN mit einer verdeckten Schicht aus 32 tanh-Einheiten, dessen Entscheidungsgrenze eine glatte nichtlineare Kurve ist. Daten generiert durch `pythondemos/binclass.py`

Der durchschnittliche 0/1-Verlust ist nützlich für die Validierung eines Klassifikators und für die Auswahl zwischen verschiedenen trainierten Klassifikatoren. Als Zielgröße für das Training ist er jedoch weniger geeignet. Tatsächlich führt die empirische Risikominimierung mit dem 0/1-Verlust zu rechnerisch schweren Optimierungsproblemen (Shalev-Shwartz and Ben-David, 2014, Kap. 12). Praktische Verfahren minimieren daher eine konvexe Ersatzfunktion des 0/1-Verlust (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Der Hinge-Verlust führt zur Stützvektormaschine (SVM), und der logistischen Verlustfunktion führt zur logistischen Regression. Abb. 3 zeigt die drei Verlustfunktionen als Funktionen des Rands $y \cdot h(\mathbf{x})$ (für den Labelraum $\mathcal{Y} = \{-1, 1\}$).

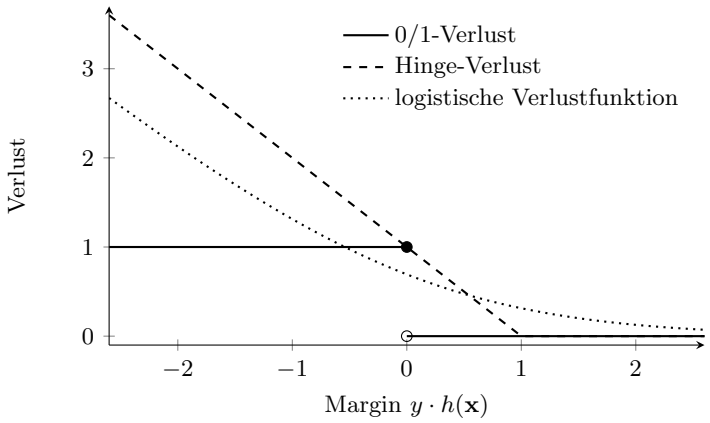


Abbildung 3: Die drei Verlustfunktionen der binären Klassifikation (mit dem Labelraum $\mathcal{Y} = \{-1, +1\}$) als Funktionen des Margins $y \cdot h(\mathbf{x})$: der 0/1-Verlust (durchgezogene Kurve; bei einem Margin von 0 nimmt er den Wert 1 an, markiert durch einen ausgefüllten Kreis, und nicht den Wert 0, markiert durch einen offenen Kreis), der Hinge-Verlust (gestrichelte Kurve) und der logistischen Verlustfunktion (gepunktete Kurve). Beide Ersatzverlustfunktionen sind konvexe und fallende Funktionen des Margins, sodass ein stärker negativer Margin mehr kostet, während der 0/1-Verlust nur einmal springt, bei einem Margin von 0

Der durchschnittliche 0/1-Verlust ist nicht nur als Zielgröße für das Training anspruchsvoll, sondern kann auch irreführend sein. Eine hohe Genauigkeit muss nicht auf einen nützlichen Klassifikator hindeuten. Von den 30 Datenpunkten des Trainingsdatensatzes und den 170 Datenpunkten des Validierungsdatensatzes zusammen gehören nur 39 zur Minderheitsklasse (Frostnächte). Ein solcher Datensatz wird als unausgewogene Daten bezeichnet, und der triviale Klassifikator, der stets den Label-Wert der Mehrheitsklasse liefert, erreicht eine Genauigkeit von 0.805.

Die Genauigkeit gibt zudem keinen Aufschluss darüber, welchen der beiden Fehler ein Klassifikator begeht, eine falsche +1 oder eine falsche −1, obwohl sich die beiden Fehler in ihren Kosten unterscheiden. Bei der Frostwarnung lässt eine falsche −1 die Blüten eines Obstbauern im Frost ungeschützt, während eine falsche +1 eine Nacht des Abdeckens oder Beheizens kostet (Snyder and de Melo-Abreu, 2005). Die beiden Fehler sind daher nicht austauschbar, und der Klassifikator wird so gewählt, dass falsche −1-Vorhersagen selten sind, zum Preis einer geringeren Genauigkeit. Die Genauigkeit gibt somit weder das Ungleichgewicht der Label-Werte noch an, welchen der beiden Fehler ein Klassifikator begeht. Solche detaillierten Informationen lassen sich aus der Konfusionsmatrix ablesen, zusammen mit den daraus berechneten Maßen Präzision, Sensitivität und F1-Score (F1-Score). Synonyme: Zweiklassen-Klassifikation.

Siehe auch: Klassifikation, Multi-Label-Klassifikation, Klassifikator, Entscheidungsgrenze, 0/1-Verlust, logistische Verlustfunktion, logistische Regressi-

on, Genauigkeit, Konfusionsmatrix, unausgewogene Daten, Baseline.

Querverweise

Literatur

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

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

3. Snyder and de Melo-Abreu (2005). Frost Protection: Fundamentals, Practice and Economics. Food and Agriculture Organization of the United Nations, Rome, Italy.

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.