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

Klassifikation ist die Aufgabe, für einen gegebenen Datenpunkt allein anhand seiner Merkmale ein diskretwertiges Label vorherzusagen. Das Label gehört zu einem endlichen Labelraum: Die binäre Klassifikation verwendet zwei Label-Werte und die Mehrklassen-Klassifikation mehr als zwei. Eine Hypothese, deren Vorhersagen Werte im endlichen Labelraum annehmen, ist ein Klassifikator; er zerlegt den Merkmalsraum in Entscheidungsregionen, die durch die Entscheidungsgrenze getrennt sind. Eine weit verbreitete Konstruktion vergleicht eine reellwertige Hypothese mit einem Schwellenwert, um das vorhergesagte Label zu erhalten. Ein natürliches Qualitätsmaß ist der 0/1-Verlust; da er weder konvex noch differenzierbar ist, minimieren Verfahren für das Training Ersatzfunktionen wie den logistischen Verlustfunktion oder den Hinge-Verlust, und der gelernte Klassifikator wird anhand der Genauigkeit und der Konfusionsmatrix auf einem Testdatensatz bewertet. Ist der Labelraum kontinuierlich, handelt es sich bei der Aufgabe um Regression.

Definition

Eine Wetterstation zeichnet jeden Tag um 06:00 und um 15:00 (UTC) die Lufttemperatur auf. Aus diesen beiden Zahlen muss entschieden werden, ob der folgende Tag Frost bringt, also ob der Messwert um 06:00 unter 0 °C fällt. Eine Bibliothek muss jeden eingehenden Text in eine der Fachkategorien *Mathematik*, *Literatur* oder *Ingenieurwesen* einordnen. Beides sind Klassifikationsprobleme: die Aufgabe, ein diskretwertiges Label y eines gegebenen Datenpunkts allein anhand seines Merkmalsvektors $\mathbf{x}$, des Vektors seiner Merkmale, vorherzusagen (Bishop, 2006; Duda et al., 2001; Hastie et al., 2009). Das Label nimmt Werte in einem endlichen Labelraum $\mathcal{Y}$ an und benennt die Kategorie des Datenpunkts. Die binäre Klassifikation verwendet einen Labelraum mit zwei Label-Werten, wie etwa $\mathcal{Y} = \{-1, 1\}$, wobei $y = 1$ einen Frosttag kennzeichnet; die Mehrklassen-Klassifikation verwendet mehr als zwei Label-Werte, wie etwa $\mathcal{Y} = \{\text{Mathematik}, \text{Literatur}, \text{Ingenieurwesen}\}$ für die Texte.

Ist der Labelraum stattdessen kontinuierlich, handelt es sich bei der Aufgabe um Regression: Die Vorhersage der morgigen Temperatur um 06:00 ist Regression, die Vorhersage, ob der morgige Tag Frost bringt, ist Klassifikation. Klassifikation unterscheidet sich auch vom Clustering: Als eine Form des

überwachten Lernens lernt die Klassifikation aus Datenpunkte, deren Labels bekannt sind, während das Clustering Datenpunkte ohne vorgegebene Labels gruppiert.

Eine Hypothese, deren Vorhersagen Werte in einem endlichen Labelraum annehmen, wird als Klassifikator bezeichnet. Ein Klassifikator zerlegt den Merkmalsraum $\mathcal{X}$ in Entscheidungsregionen – eine Region pro Label-Wert –, die durch die Entscheidungsgrenze getrennt sind. Alle Merkmalsvektoren in derselben Entscheidungsregion erhalten dasselbe vorhergesagte Label, sodass die Entscheidungsregionen den Klassifikator vollständig bestimmen (Abb. 1).

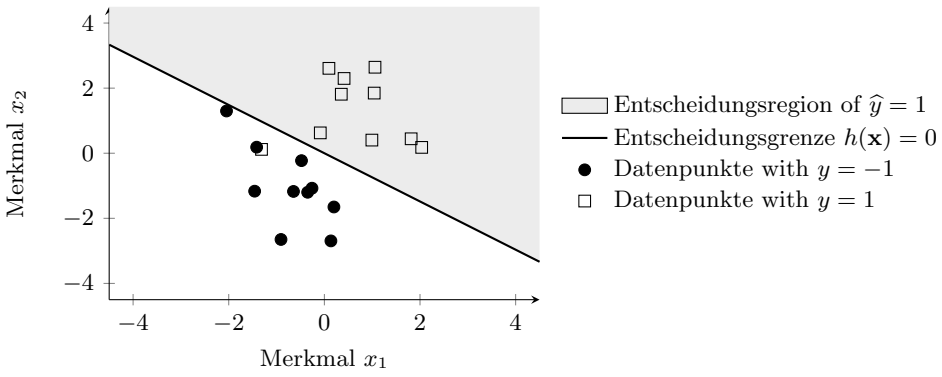

Abbildung 1: Binäre Klassifikation mit zwei Merkmale $x_1, x_2 \in \mathbb{R}$. Die ausgefüllten Kreise sind Datenpunkte mit $y = -1$, und die offenen Quadrate sind Datenpunkte mit $y = 1$. Die durchgezogene Linie ist die Entscheidungsgrenze; sie teilt $\mathcal{X} = \mathbb{R}^2$ in eine nicht schattierte Entscheidungsregion mit $\widehat{y} = -1$ und eine schattierte Region mit $\widehat{y} = 1$. Ein offenes Quadrat liegt in der nicht schattierten Region, in der die Vorhersage $\widehat{y} = -1$ lautet, sodass dieser Datenpunkt falsch klassifiziert wird. Daten erzeugt mit `pythondemos/classification.py`

Eine weit verbreitete Konstruktion eines Klassifikators erfolgt in zwei Schritten. Zuerst quantifiziert eine reellwertige Hypothese $h : \mathcal{X} \to \mathbb{R}$ die Konfidenz für einen bestimmten Label-Wert. Zweitens wird die Konfidenz $h(\mathbf{x})$ mit einem Schwellenwert verglichen, um die Vorhersage $\widehat{y} \in \mathcal{Y}$ zu erhalten. Die logistische Regression verwendet für den Labelraum $\mathcal{Y} = \{-1, 1\}$ eine lineare Abbildung $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ als Konfidenz für den Label-Wert 1. Diese Konfidenz wird mit dem Schwellenwert 0 verglichen,

$$\widehat{y} := \begin{cases} 1 & \text{falls } h(\mathbf{x}) \geq 0, \\ -1 & \text{andernfalls.} \end{cases}$$

Die reellwertige Hypothese h ist nicht selbst der Klassifikator: Der Klassifikator ist die schwellenwertbasierte Abbildung $\mathbf{x} \mapsto \widehat{y}$, und h als Klassifikator zu bezeichnen, ist sprachlich etwas ungenau. Die beiden Entscheidungsregionen sind

dann die Halbräume auf beiden Seiten der Hyperebene $\mathbf{w}^\top \mathbf{x} = 0$, wie in Abb. 1 dargestellt.

Ein natürliches Qualitätsmaß für einen Klassifikator ist der 0/1-Verlust: Sein Durchschnitt über einen Datensatz ist der Anteil falsch klassifizierter Datenpunkte, und eins minus diesen Anteil ist die Genauigkeit. Selbst der beste Klassifikator kann nicht immer null Fehler erreichen: Wenn Datenpunkte mit demselben Merkmalsvektor unterschiedliche Labels tragen, sind einige Fehlklassifikationen unvermeidbar, und das kleinste erreichbare Risiko ist das Bayes-Risiko (Hastie et al., 2009, Abschn. 2.4, Gl. (2.23)).

Der 0/1-Verlust wird selten als Zielfunktion für das Training verwendet: Er nimmt nur die Werte 0 und 1 an, ist also stückweise konstant, als Funktion der Modellparameter weder konvex noch differenzierbar, und seine Minimierung ist berechnungstechnisch schwer (Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Praktische Verfahren minimieren daher den Durchschnitt eines Ersatz-Verlusts über den Trainingsdatensatz: eine Verlustfunktion, die den 0/1-Verlust approximiert und zugleich günstigere Eigenschaften wie Konvexität oder Differenzierbarkeit aufweist. Die logistische Regression verwendet den logistischen Verlustfunktion, der konvex und differenzierbar ist (Bishop, 2006, Abschn. 4.3.3; Shalev-Shwartz and Ben-David, 2014, Abschn. 9.3); die Stützvektormaschine (SVM) verwendet den Hinge-Verlust, der konvex, aber nicht differenzierbar ist (Boser et al., 1992; Cortes and Vapnik, 1995; Shalev-Shwartz and Ben-David, 2014, Abschn. 12.3). Abb. 2 vergleicht die drei als Funktionen des Margins $y \cdot h(\mathbf{x})$. Der gelernte Klassifikator wird anschließend anhand der Genauigkeit und der Konfusionsmatrix auf einem Testdatensatz bewertet.

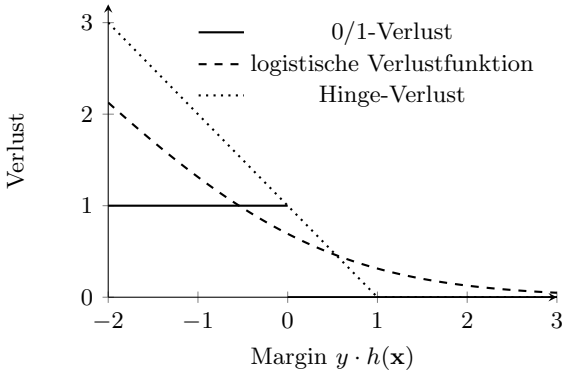

Abbildung 2: Der 0/1-Verlust (durchgezogene Linie), der logistischen Verlustfunktion (gestrichelte Kurve) und der Hinge-Verlust (gepunktete Linie), dargestellt als Funktionen des Margins $y \cdot h(\mathbf{x})$ für $\mathcal{Y} = \{-1, 1\}$. Der 0/1-Verlust beträgt 1 für einen nichtpositiven Rand und 0 für einen positiven Rand. Der logistischen Verlustfunktion ist konvex und differenzierbar. Der Hinge-Verlust ist konvex, doch sein Knick bei Rand 1 zeigt, dass er dort nicht differenzierbar ist

Siehe auch: Label, Labelraum, Klassifikator, binäre Klassifikation, Entscheidungsgrenze, Entscheidungsregion, 0/1-Verlust, logistische Verlustfunktion, Hinge-Verlust, logistische Regression, SVM, Genauigkeit, Konfusionsmatrix, Regression, überwachtes Lernen, Clustering, Bayes-Risiko.

Querverweise

Literatur

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

2. Boser et al. (1992). A Training Algorithm for Optimal Margin Classifiers. In: Proc. 5th Annual Workshop on Computational Learning Theory (COLT), pp. 144–152. ACM.

3. Duda et al. (2001). Pattern Classification. 2nd ed., Wiley, New York, NY, USA.

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

5. Cortes and Vapnik (1995). Support-vector networks. Machine learning, 20(3), pp. 273–297.

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.