linearer Klassifikator

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

Ein linearer Klassifikator ist ein Klassifikator, dessen Entscheidungsregionen durch Hyperebenen getrennt sind. Jeder lineare Klassifikator entsteht durch den Vergleich des Werts einer linearen Hypothese-Abbildung mit einem Schwellenwert. Für binäre Klassifikation kann das Vorzeichen des Werts zur Bestimmung des vorhergesagten Label verwendet werden, und der Absolutwert geteilt durch die Euclidean norm des Gewicht-Vektor ist die Abstand des Datenpunkt von der Entscheidungsgrenze. Die Modellparameter eines linearen Klassifikators sind der Gewicht-Vektor der zugrundeliegenden Hypothese-Abbildung und der Offset. Sie werden durch empirische Risikominimierung (ERM) mit verschiedenen Verlustfunktionen gelernt: logistische Regression verwendet die logistische Verlustfunktion und die Stützvektormaschine (SVM) die hinge loss. Für die Mehrklassen-Klassifikation ist ein linearer Klassifikator durch mehrere lineares Modell-Hypothese-Abbildungen charakterisiert, eine pro Label-Wert, und die resultierenden Entscheidungsregionen sind Schnitte von Halbräume.

Definition

Eine Luftaufnahme einer Weinbauregion wird in kleine Bereiche unterteilt, und jeder Bereich soll als Weinberg oder Nicht-Weinberg klassifiziert werden. Jeder Bereich ist ein Datenpunkt, der durch zwei Merkmale aus dem Foto charakterisiert ist (Abb. 1). Das Label jedes Bereichs gibt an, ob der Bereich zu einem Weinberg gehört ($y = 1$) oder nicht ($y = -1$). Die gelabelten Bereiche bilden das Trainingsdatensatz. Ein linearer Klassifikator trennt die beiden Klassen durch eine Gerade, seine Entscheidungsgrenze, die durch einen Normalenvektor festgelegt ist, und klassifiziert einen neuen Bereich anhand seiner vorzeichenbehafteten Abstand zu dieser Geraden.

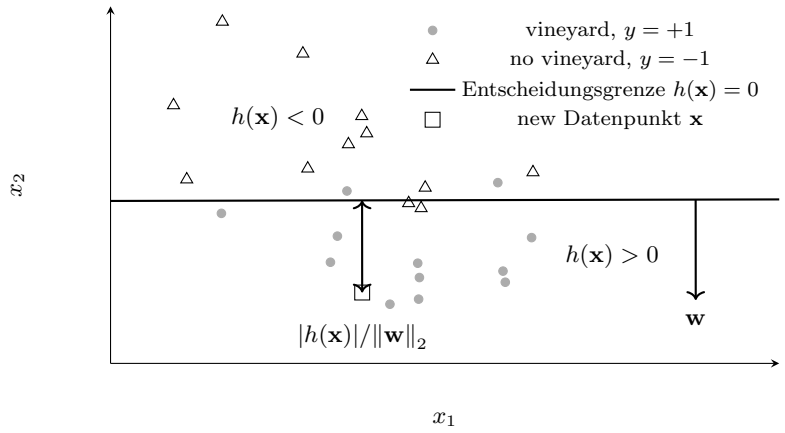

Abbildung 1: Scatterplot von Datenpunkte, jeweils charakterisiert durch zwei Merkmale und ein binäres Label. Die durchgezogene Linie ist die Entscheidungsgrenze $h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b = 0$ eines aus diesen Datenpunkte gelernten linearen Klassifikators, und der einseitige Pfeil ist sein Normalenvektor $\mathbf{w}$. Der doppelseitige Pfeil spannt $|h(\mathbf{x})|/\|\mathbf{w}\|_2$ auf, die Abstand von der Entscheidungsgrenze eines neuen Datenpunkt $\mathbf{x}$ (offenes Quadrat), der nicht im Trainingsdatensatz enthalten ist. Daten erzeugt mit `pythondemos/linclass.py`

Ein linearer Klassifikator ist ein Klassifikator, dessen Entscheidungsregionen durch Hyperebenen in $\mathbb{R}^d$ getrennt sind (Bishop, 2006, Sect. 4.1.1; Shalev-Shwartz and Ben-David, 2014, Sect. 9.1). Für binäre Klassifikation mit Labelraum $\mathcal{Y} = \{-1, +1\}$ kann jeder lineare Klassifikator als Schwellwertregel $\widehat{y} = \text{sign}(h(\mathbf{x}))$ mit der Hypothese

$$h(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + b.$$

Die Hypothese-Abbildung ist durch den Gewicht-Vektor $\mathbf{w} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ und den Offset $b \in \mathbb{R}$ parametrisiert, der auch als Schwellenwert oder Bias bezeichnet wird. Ein Datenpunkt mit Merkmalsvektor $\mathbf{x}$ wird als $\widehat{y} = 1$ klassifiziert, falls $h(\mathbf{x}) \geq 0$, und als $\widehat{y} = -1$ sonst. Die resultierende Entscheidungsgrenze ist die Hyperebene

$$\{\mathbf{x} \in \mathbb{R}^d : \mathbf{w}^\top \mathbf{x} + b = 0\},$$

mit Normalenvektor $\mathbf{w}$. Diese Grenze trennt zwei Halbräume, nämlich die Entscheidungsregionen $\mathcal{R}_1 = \{\mathbf{x} : h(\mathbf{x}) \geq 0\}$ und $\mathcal{R}_{-1} = \{\mathbf{x} : h(\mathbf{x}) < 0\}$. Die Merkmalsvektoren auf der Entscheidungsgrenze, mit $h(\mathbf{x}) = 0$, werden konventionsgemäß $\mathcal{R}_1$ zugeordnet.

Betrachte einen Merkmalsvektor $\mathbf{x}$ und einen beliebigen Vektor $\mathbf{x}'$ auf der Entscheidungsgrenze, $h(\mathbf{x}') = 0$. Subtrahiert man $\mathbf{w}^\top \mathbf{x}' + b = 0$ von $h(\mathbf{x})$, ergibt sich

$$h(\mathbf{x}) = \mathbf{w}^\top \big(\mathbf{x} - \mathbf{x}'\big).$$

Die Cauchy-Schwarz inequality beschränkt die rechte Seite im Betrag durch $\|\mathbf{w}\|_2\|\mathbf{x}-\mathbf{x}'\|_2$ (Axler, 2024, Thm. 6.14). Jeder Vektor der Entscheidungsgrenze ist daher in einer Abstand von mindestens $|h(\mathbf{x})|/\|\mathbf{w}\|_2$ von $\mathbf{x}$. Diese Abstand wird durch die orthogonale Projektion

$$\mathbf{x}' = \mathbf{x} - \frac{h(\mathbf{x})}{\|\mathbf{w}\|_2^2}\,\mathbf{w}.$$

erreicht. Die Abstand von $\mathbf{x}$ zur Entscheidungsgrenze beträgt somit $|h(\mathbf{x})|/\|\mathbf{w}\|_2$ (Bishop, 2006, Sect. 4.1.1; Hastie et al., 2009, Sect. 4.5, Eq. (4.40)). Abb. 1 illustriert die Abstand zur Entscheidungsgrenze für einen Datenpunkt.

Der Wert $|h(\mathbf{x})|$ ist ein Maß für die Störungsrobustheit am Merkmalsvektor $\mathbf{x}$. Eine Störung von $\mathbf{x}$ um einen Vektor $\boldsymbol{\delta}$ ändert den Hypothese-Wert um $\mathbf{w}^\top\boldsymbol{\delta}$, was durch die Cauchy-Schwarz inequality im Betrag durch $\|\mathbf{w}\|_2\|\boldsymbol{\delta}\|_2$ beschränkt wird. Das Vorzeichen von $h(\mathbf{x})$, und damit die Vorhersage, bleibt daher bei jeder Störung mit $\|\boldsymbol{\delta}\|_2 < |h(\mathbf{x})|/\|\mathbf{w}\|_2$ unverändert.

Multipliziert man $h(\mathbf{x})$ mit dem Label y, ergibt sich der Margin. Ein positiver Margin bedeutet, dass der Datenpunkt korrekt klassifiziert ist, ein negativer Margin, dass er fehlklassifiziert ist, und ein Null- Margin, dass er auf der Entscheidungsgrenze liegt, wobei nach obiger Regel $\widehat{y} = 1$ zugeordnet wird.

Die Modellparameter $(\mathbf{w}, b)$ werden aus einem Trainingsdatensatz $\mathcal{D}^{(\text{train})} = \{(\mathbf{x}^{(r)}, y^{(r)})\}_{r=1}^m$ durch empirische Risikominimierung (ERM) gelernt. Die Verwendung des $0/1$ loss führt zu einer Zielfunktion, die stückweise konstant ist, und deren Minimierung ist im Allgemeinen rechnerisch schwierig (Shalev-Shwartz and Ben-David, 2014, Sect. 9.1). Praktische Verfahren ersetzen den $0/1$ loss durch einen konvex Verlust des Margin: logistische Regression verwendet die logistische Verlustfunktion und die Stützvektormaschine (SVM) die hinge loss. Für ein linear trennbares Trainingsdatensatz lernt die SVM mit hinreichend kleinem Regularisierung-Parameter α, dem Faktor, der ihren Strafterm skaliert, den linearen Klassifikator mit größtem kleinstem Margin (Shalev-Shwartz and Ben-David, 2014, Sect. 15.1, Eq. (15.1)). Beide lernen die Modellparameter eines linearen Klassifikators durch Minimierung einer konvex Zielfunktion: Gradientenabstieg minimiert die smooth Zielfunktion der logistische Regression, und subgradient descent die non-smooth Zielfunktion der SVM (Murphy, 2012, Sect. 8.3.2; Shalev-Shwartz and Ben-David, 2014, Sect. 14.2.3, Example 14.2). Die squared error loss ist ebenfalls eine Funktion des Margin, denn $(y - h(\mathbf{x}))^2 = (1 - y\,h(\mathbf{x}))^2$ für $y \in \{-1, +1\}$, sodass lineare Regression, Ridge-Regression und least absolute shrinkage and selection operator (Lasso) auf die Labels ± 1 ebenfalls lineare Klassifikatoren lernen (Bishop, 2006, Sect. 4.1.3).

Für Mehrklassenprobleme mit Labelraum $\mathcal{Y} = \{1, \ldots, K\}$ und $K > 2$ verwendet ein linearer Klassifikator je einen lineares Modell-Hypothese $g_c(\mathbf{x}) = \left(\mathbf{w}^{(c)}\right)^\top\mathbf{x} + b^{(c)}$ pro Label-Wert $c \in \mathcal{Y}$ und sagt den Wert mit dem größten Score vorher, $\widehat{y} = \arg\max_{c\in\mathcal{Y}} g_c(\mathbf{x})$. Die Entscheidungsregion $\mathcal{R}_c$ ist dann ein Schnitt von $K-1$ Halbräume und damit eine konvex Menge. Die Grenze zwischen zwei Entscheidungsregionen $\mathcal{R}_c$ und $\mathcal{R}_{c'}$ liegt in der Hyperebene, auf der die beiden Scores übereinstimmen (Bishop, 2006, Sect. 4.1.2, Eq. (4.10)). Die resultierende

Partition des Merkmalsraum ist in Abb. 2 für einen linearen Klassifikator mit $K = 3$ Klassen dargestellt.

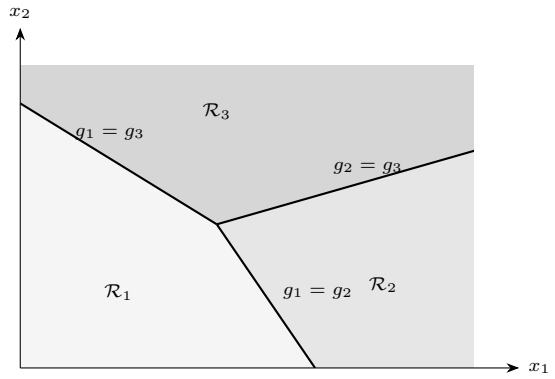

Abbildung 2: Ein linearer Klassifikator für $K = 3$ Label-Werte zerlegt den Merkmalsraum $\mathbb{R}^2$ in drei Entscheidungsregionen $\mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3$. Jede Grenze liegt in der Hyperebene, auf der die beiden konkurrierenden Scores übereinstimmen, und jede Entscheidungsregion ist ein Schnitt von Halbräume und daher konvex

Ein linearer Klassifikator kann verallgemeinert werden, um nichtlineare Entscheidungsgrenzen zuzulassen, indem eine Merkmalstransformation verwendet wird. Wird zunächst eine Merkmalstransformation $\phi : \mathbb{R}^d \to \mathbb{R}^{d'}$ angewendet und anschließend ein linearer Klassifikator auf $\phi(\mathbf{x})$, so ergibt sich eine Entscheidungsgrenze, die ein Hyperebene in $\mathbb{R}^{d'}$, aber nichtlinear im ursprünglichen Merkmalsraum ist. Eine Kernel-Methode konstruiert eine solche Merkmalstransformation aus einem Kernel, und ein tiefes Netz nutzt seine ersten Schichten als trainierbare Merkmalstransformation (Bishop and Bishop, 2024, Sect. 6.3.3; Schölkopf and Smola, 2002, Sect. 2.2).

Siehe auch: Klassifikator, binäre Klassifikation, Entscheidungsgrenze, Entscheidungsregion, Hyperebene, Halbraum, lineares Modell, Margin, logistische Regression, SVM, ERM, 0/1 loss, Ridge-Regression, Lasso, Merkmalstransformation.

Querverweise

Literatur

1. Axler (2024). Linear Algebra Done Right. 4th ed., Springer, Cham, Switzerland.

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

3. Bishop and Bishop (2024). Deep Learning: Foundations and Concepts. Springer Nature, Cham, Switzerland.

4. Schölkopf and Smola (2002). Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge, MA, USA.

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

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

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