Entscheidungsbaum

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 Entscheidungsbaum ist eine Flussdiagramm-ähnliche Darstellung einer Hypotheseabbildung: Ein Wurzelknoten empfängt den Merkmalsvektor eines Datenpunkts und führt einen grundlegenden Test darauf aus. Das Testergebnis wählt einen Kindknoten aus, der einen weiteren Test anwendet, und die Weiterleitung wird fortgesetzt, bis der Merkmalsvektor einen Blattknoten erreicht. Ein Blattknoten hat keine Kinder und repräsentiert einen spezifischen Funktionswert, d.h. eine Vorhersage. Der Baum partitioniert den Merkmalsraum in Entscheidungsregionen, eine pro Blattknoten, auf denen die dargestellte Hypothese konstant ist. Eine maximale Tiefe und eine Familie von Knotentests definieren das Modell, den Hypothesenraum aller solcher Bäume, der keinen Vektor von Modellparameter fester Länge besitzt und daher gierig statt mittels Gradient-Methoden durchsucht wird. Entscheidungsbäume dienen als Basismodelle des Random Forest und der gradient-boosted decision tree (GBDT)-Methoden. Kleine Entscheidungsbäume führen zu interpretierbares maschinelles Lernen, wenn der Benutzer die an ihren Knoten ausgeführten Tests nachvollziehen kann.

Definition

Die Kreditgenehmigung einer Bank kann als eine Kette einfacher Tests der Merkmale des Antragstellers geschrieben werden, die in einer „genehmigen“- oder „ablehnen“-Vorhersage endet. Zum Beispiel: Liegt das Einkommen über einem Schwellenwert? Ist die Kreditwürdigkeit ausreichend hoch? Ein Entscheidungsbaum ist eine solche Kette in allgemeiner Form: eine Flussdiagramm-ähnliche Darstellung einer Hypotheseabbildung h.

Formaler ausgedrückt ist ein Entscheidungsbaum ein directed acyclic graph (DAG), dessen Wurzelknoten den Merkmalsvektor $\mathbf{x}$ eines Datenpunkts empfängt. Der Wurzelknoten leitet den Datenpunkt dann basierend auf einem elementaren Test der Merkmale $\mathbf{x}$ an einen seiner Kindknoten weiter. Wenn der empfangende Kindknoten kein Blattknoten ist, d.h. selbst Kindknoten hat, repräsentiert er einen weiteren Test. Basierend auf dem Testergebnis wird der Datenpunkt an einen seiner Nachkommen weitergeleitet. Dieses Testen und Weiterleiten des Datenpunkts wird fortgesetzt, bis der Datenpunkt in einem Blattknoten ohne Kinder landet, der die vom Baum für diesen Datenpunkt

zurückgegebene Vorhersage $\widehat{y} = h(\mathbf{x})$ trägt (Abb. 1). Die Blattknoten partitionieren den Merkmalsraum in Entscheidungsregionen; in Abb. 1 vergleicht jeder Test einen Eintrag von $\mathbf{x}$ mit einem Schwellenwert, sodass die Entscheidungsregionen Rechtecke sind (Hastie et al., 2009, Abschn. 9.2.1). Neben jedem Knoten hebt Abb. 1 den Teil des Merkmalsraum hervor, der die Merkmalsvektoren enthält, die diesen Knoten erreichen.

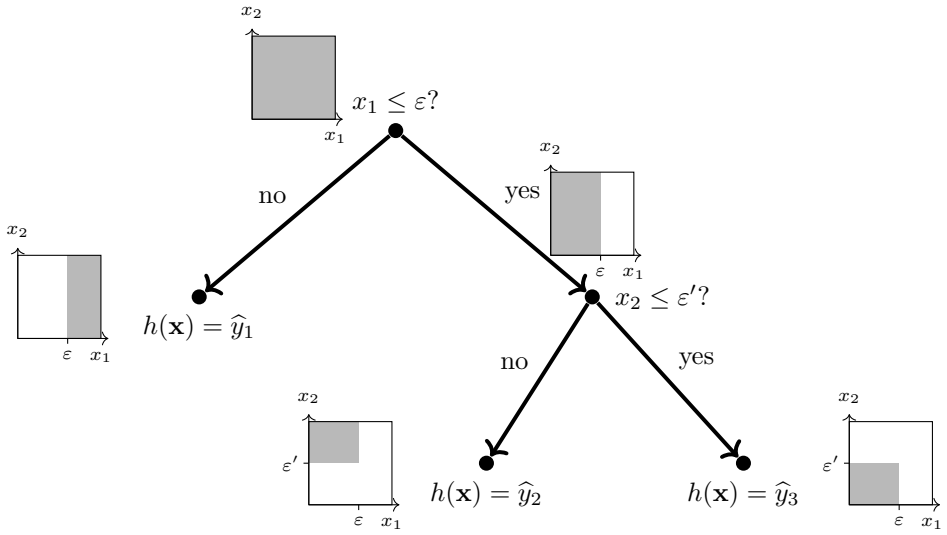


Abbildung 1: Entscheidungsbaum als flussdiagrammähnliche Darstellung einer stückweise konstanten Hypothese $h : \mathcal{X} \to \mathbb{R}$ für Merkmalsvektoren $\mathbf{x} = (x_1, x_2)^\top$. Jeder Test vergleicht ein einzelnes Merkmal mit einem der Schwellenwerte $\varepsilon, \varepsilon' > 0$, die an den Achsen jedes Quadrats markiert sind. Neben jedem Knoten enthält der schattierte Teil des Merkmalsraum (dargestellt als Quadrat mit den Achsen x_1 und x_2) die Merkmalsvektoren, die diesen Knoten erreichen. Die schattierten Teile an den Blattknoten sind die Entscheidungsregionen $\mathcal{R}_{\widehat{y}} := \{\mathbf{x} \in \mathcal{X} : h(\mathbf{x}) = \widehat{y}\}$: Rechtecke, da jeder Test ein einzelnes Merkmal beinhaltet

Die Tiefe eines Entscheidungsbaums ist die Anzahl der Tests auf dem längsten Pfad vom Wurzelknoten zu einem Blattknoten; der Baum in Abb. 1 hat Tiefe zwei. Das Festlegen einer maximalen Tiefe D und einer Familie zulässiger Tests definiert das Modell: den Hypothesenraum $\mathcal{H}$ aller Entscheidungsbäume mit einer Tiefe von höchstens D, deren Knotentests zur Familie gehören, zum Beispiel alle Schwellenwerttests $x_j \leq \varepsilon$ mit $j \in \{1, \ldots, d\}$ und $\varepsilon \in \mathbb{R}$ (Shalev-Shwartz and Ben-David, 2014, Abschn. 18.1). Ein Baum in $\mathcal{H}$ wird durch seine Form, durch den Index j des Merkmals und den Schwellenwert ε, den jeder Testknoten vergleicht, sowie durch den Wert jedes Blattknotens spezifiziert.

Dieser Hypothesenraum hat keine natürliche Parametrisierung. Die Form des Baumes und der Index j des an jedem Knoten getesteten Merkmals sind diskrete Entscheidungen, und wie viele davon ein Baum hat, hängt von seiner Form ab, sodass $\mathcal{H}$ nicht das Bild eines Vektors von Modellparameter fester Länge ist: Es gibt keinen parameter space $\mathbb{R}^n$, auf dem der Trainingsfehler eine Funktion ist. Die Schwellenwerte bilden einen Vektor, sobald die Form und die getesteten Merkmale festgelegt sind, aber der Trainingsfehler ist in jedem von ihnen stückweise konstant, da ein Datenpunkt seinen Kindknoten nur ändert, wenn der Schwellenwert seinen Merkmal-Wert überschreitet, sodass der Gradient überall dort, wo er existiert, Null ist. Empirische Risikominimierung (ERM) über $\mathcal{H}$ kann daher keine Gradientenabstieg verwenden, und eine Fixpunkt-Betrachtung der Training-Iteration ist nicht anwendbar: Eine Iteration fügt einen Knoten hinzu, sie aktualisiert keinen Vektor. Das Minimieren einer Schranke, die dem Trainingsfehler eine Größenstrafe hinzufügt, ist rechnerisch aufwendig, und die unten beschriebenen Methoden durchsuchen $\mathcal{H}$ gierig (Shalev-Shwartz and Ben-David, 2014, Abschn. 18.2).

Das Training eines Entscheidungsbaums besteht darin, einen Blattknoten durch Anhängen eines weiteren Entscheidungsbaums zu erweitern. Im einfachsten Fall wird ein Blattknoten durch einen Testknoten ersetzt, dessen Kinder neue Blattknoten sind (Abb. 2).

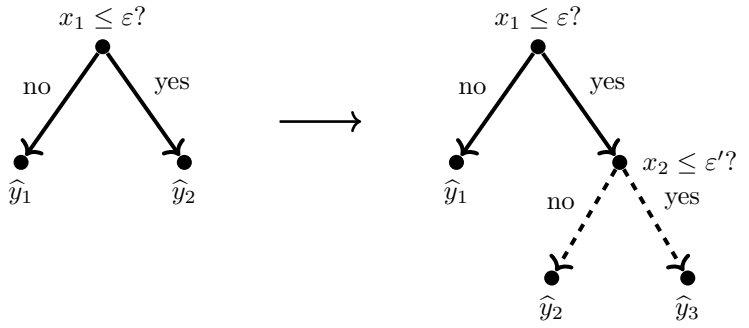


Abbildung 2: Das Training lässt einen Entscheidungsbaum wachsen. Der Baum rechts wird aus dem Baum links erhalten, indem der Blattknoten mit Vorhersage $\widehat{y}_2$ durch einen Testknoten ersetzt wird, dessen Kinder zwei neue Blattknoten sind (gestrichelte Kanten)

Weit verbreitete Training-Methoden für Entscheidungsbäume umfassen Klassifikations- und Regressionsbäume (CART) (Breiman et al., 1984), Iterative Dichotomiser 3 (ID3) (Quinlan, 1986) und C4.5 (Quinlan, 1993). Ausgehend von einem einzelnen Blattknoten ersetzt CART wiederholt einen Blattknoten durch einen Testknoten. Ein Unreinheitsmaß weist einer Menge von Labels eine nicht-negative Zahl zu: Es ist Null, wenn die Labels übereinstimmen, und es wächst, wenn sie sich ausbreiten. Unter den Kandidatentests, die ein einzelnes Merkmal mit einem Schwellenwert vergleichen, bewertet CART jeden Test nach

der Unreinheit der beiden Kinder, gewichtet nach dem Anteil der an jedes von ihnen weitergeleiteten Datenpunkte, und wählt gierig den Test mit der kleinsten Bewertung, d.h. mit der größten Reduktion gegenüber der Unreinheit des Elternknotens. Gängige Unreinheitsmaße sind der Gini-Index und die entropy für die Klassifikation sowie der quadratische Fehler um das Mittelwert Label für die regression (Hastie et al., 2009, Abschn. 9.2.2 und 9.2.3). Das Wachstum stoppt nur, wenn ein Blattknoten zu wenige Datenpunkte enthält, nicht wenn kein Kandidatentest die Unreinheit stark reduziert, da ein Test, der die Unreinheit wenig reduziert, einen Test mit einer großen Unreinheitsreduktion darunter ermöglichen kann (Hastie et al., 2009, Abschn. 9.2.2). Der große Baum wird dann durch Kosten-Komplexitäts-Pruning zurückgeschnitten, das die Unterbäume entfernt, deren Unreinheitsreduktion kleiner ist als die Strafe für die von ihnen hinzugefügten Blattknoten (Hastie et al., 2009, Abschn. 9.2.2).

Abb. 3 zeigt als scatterplot das Ergebnis dieses Trainings auf realen Daten. Jeder Tag des Jahres 2024 an der Wetterstation Krems an der Donau (Österreich), der beide Messwerte aufweist, ist ein Datenpunkt, wobei die Lufttemperatur um 06:00 (UTC) als sein Merkmal und die Temperatur um 15:00 (UTC) desselben Tages als sein Label dient. Der gelernte Baum der Tiefe 2 ist eine stückweise konstante Hypothese mit vier Stücken, eines pro Blattknoten, wobei jedes dem Mittelwert Label der dorthin geleiteten Tage entspricht. Der Baum wurde bis zur Tiefe zwei gewachsen, anstatt bis die oben genannte Abbruchregel angewendet wurde, sodass jedes seiner vier Stücke im Bild sichtbar ist.

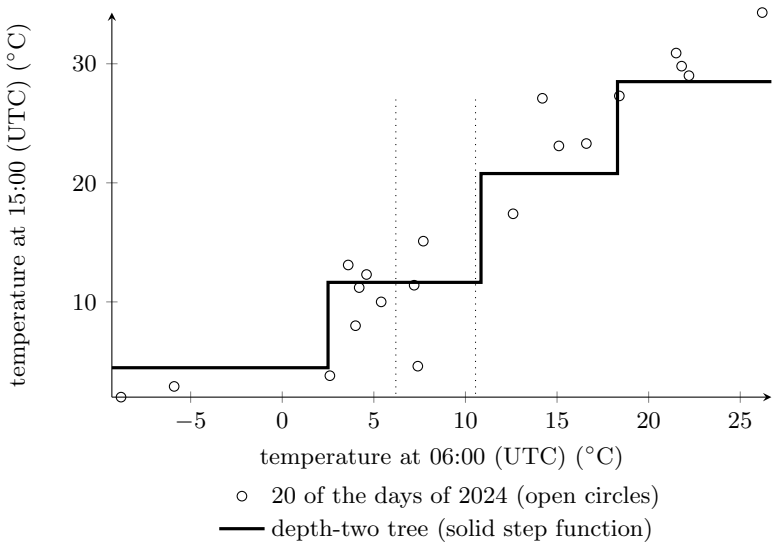


Abbildung 3: Jeder der 365 Tage des Jahres 2024 in Krems an der Donau (Österreich), der beide Messwerte aufweist, ist ein Datenpunkt: sein Merkmal ist die Lufttemperatur um 06:00 (UTC) (horizontale Achse) und sein Label die Temperatur um 15:00 (UTC) desselben Tages (vertikale Achse). Die offenen Kreise stellen 20 davon dar, jeden 19. Tag. Die dicke durchgezogene Treppenfunktion ist die stückweise konstante Hypothese $\widehat{h}$, die aus allen 365 Tagen von einem regressionsbaum der Tiefe zwei gelernt wurde. Die drei Schwellenwerte seiner Testknoten, bei 2.50 °C, 10.85 °C und 18.30 °C (gestrichelte Vertikalen), sind die Positionen der drei Sprünge und teilen die Merkmal-Achse in vier Intervalle. Auf jedem Intervall ist die Treppenfunktion flach beim Mittelwert Label der Datenpunkte, die zu diesem Blattknoten geleitet wurden. Daten generiert von `pythondemos/decisiontree.py`

Entscheidungsbäume dienen als Basismodelle des Random Forest und der gradient-boosted decision tree (GBDT)-Methoden (Hastie et al., 2009, Kap. 10 und 15). Ein Entscheidungsbaum mit einer geringen Anzahl von Knoten wird oft als interpretierbar angesehen: Seine Berechnung kann nachvollzogen werden, indem man den Datenpunkt vom Wurzelknoten zu einem Blattknoten verfolgt (Molnar, 2025). Kleine Entscheidungsbäume sind daher eine gängige Wahl in der interpretierbares maschinelles Lernen.

Siehe auch: Entscheidungsregion, Entscheidungsregel, Klassifikation, Random Forest, GBDT, Merkmalswichtigkeit, interpretierbares maschinelles Lernen.

Querverweise

Literatur

1. Breiman et al. (1984). Classification and Regression Trees. Wadsworth, Belmont, CA.
2. Molnar (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 3rd ed., Ebook.
3. Quinlan (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann, San Mateo, CA.
4. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.
5. Quinlan (1986). Induction of decision trees. Machine Learning, 1(1), pp. 81–106.
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.