Random Forest

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 Random Forest ist die Instanz von Bagging, dessen Basislerner Entscheidungsbäume lernen. Jeder Entscheidungsbaum wird auf zufällig ausgewählten Datenpunkte des Trainingsdatensatz trainiert, und jeder Basislerner verwendet auch nur eine zufällig ausgewählte Untermenge der Merkmale. Die Vorhersagen der Basislerner werden durch Mehrheitsentscheid bei der Klassifikation oder durch Mittelwertbildung bei der regression aggregiert. Die zufällige Auswahl von Datenpunkte und von Merkmale zielt auf eine Dekorrelation ab, sodass die Aggregation die Varianz der zurückgegebenen Vorhersage reduziert. Die Aggregation lässt den bias unverändert, und das Hinzufügen von Basislerner führt nicht zu Überanpassung. Ein Random Forest liefert auch einen Merkmalswichtigkeit-Wert pro Merkmal: den Anteil der gesamten Splits, die dieses Merkmal verwenden.

Definition

Eine Bank muss für jeden Kreditantrag entscheiden, ob der Antragsteller kreditwürdig ist. Ein Random Forest antwortet mit der Mehrheitsentscheidung vieler verschiedener Entscheidungsbäume. Ein Random Forest ist die Instanz von Bagging, dessen Basislerner Entscheidungsbäume lernen. Wie bei Bagging wird jeder Entscheidungsbaum auf zufällig ausgewählten Datenpunkte des Trainingsdatensatz trainiert. Neben der zufälligen Auswahl von Datenpunkte verwendet jeder Basislerner auch nur eine zufällig ausgewählte Untermenge der Merkmale der ursprünglichen Datenpunkte (Breiman, 2001). Die Vorhersagen der Basislerner werden durch Mehrheitsentscheid bei der Klassifikation oder durch Mittelwertbildung bei der regression (Abb. 2). Das scatterplot von Abb. 1 zeigt ein regression-Beispiel: Jeder Datenpunkt ist einer der 366 Tage des Jahres 2020 an der Wetterstation Krems an der Donau, sein Merkmal die minimale und sein Label die maximale Lufttemperatur des Tages. Drei Entscheidungsbäume, jeder auf seiner eigenen zufälligen Auswahl der Tage trainiert, ergeben drei verschiedene stückweise konstante Hypothesen; der Random Forest liefert deren Durchschnitt.

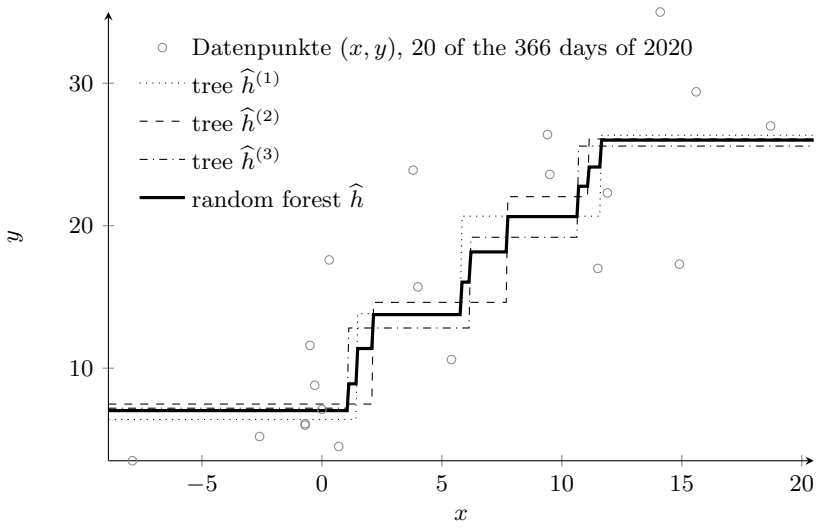


Abbildung 1: Scatterplot von 20 der 366 Tage des Jahres 2020 an der Wetterstation Krems an der Donau (jeder 19. Tag), mit der Minimaltemperatur eines Tages als Merkmal x und seiner Maximaltemperatur als Label y. Drei Entscheidungsbäume $\widehat{h}^{(1)}, \widehat{h}^{(2)}, \widehat{h}^{(3)}$, jeder auf seiner eigenen zufälligen Auswahl der Tage trainiert (gepunktete, gestrichelte und strichpunktierte Stufenkurven). Die dicke durchgezogene Stufenkurve ist der Random Forest $\widehat{h}$, der den Durchschnitt der drei liefert. Daten generiert von `pythondemos/randomforest.py`

Diese Konstruktion funktioniert, weil das Mitteln vieler Vorhersage mit hoher Varianz eine mit geringerer Varianz ergibt, vorausgesetzt, die Fehler der einzelnen Vorhersagen sind schwach korreliert. Ein tiefer Entscheidungsbaum ist hier als Basislerner gut geeignet: Er kann das Trainingsdatensatz eng anpassen (geringer bias), aber seine Vorhersage variiert stark mit dem Trainingsdatensatz (große Varianz). Das Mitteln reduziert diese Varianz und lässt den bias unverändert, da die Bäume identisch verteilt sind (Hastie et al., 2009, Sect. 15.2). Das Hinzufügen von Bäumen führt nicht zu Überanpassung: Mit zunehmender Anzahl konvergiert die gemittelte Vorhersage gegen ihren Erwartungswert über die Zufallsauswahlen, und nur dieser Grenzwert kann überanpassen (Hastie et al., 2009, Sect. 15.3.4). Betrachten wir M identisch verteilte Vorhersagen, jede mit Varianz σ^2 und mit paarweiser Korrelation ρ. Die gemittelte Vorhersage hat die Varianz

$$\rho\sigma^2 + \frac{1-\rho}{M}\,\sigma^2,$$

die mit der Anzahl M der Bäume gegen die untere Grenze $\rho\sigma^2$ abnimmt, die durch die Korrelation festgelegt ist (Hastie et al., 2009, Gl. (15.1)). Die Zufallsauswahlen werden unabhängig voneinander gezogen, sobald das Trainingsdatensatz festgelegt ist, was die Vorhersagen identisch verteilt macht. Sie sind nicht unabhängig, wenn das zugrunde liegende Datensatz selbst eine Zufalls-

variable (ZV) ist: Jeder Entscheidungsbaum wird aus demselben Trainingsdatensatz gezogen, und diese gemeinsame Abhängigkeit führt im Allgemeinen zu einer Korrelation ρ ungleich Null (Hastie et al., 2009, Sect. 15.4.1, Gl. (15.6)).

Die zufällige Auswahl von Datenpunkte und die zufällige Auswahl von Merkmale zielen beide auf eine Dekorrelation der Vorhersagen ab, die die Basislerner für einen Merkmalsvektor $\mathbf{x}$ liefern (Hastie et al., 2009, Sect. 15.2). Während das Zurückhalten von Merkmale die Korrelation ρ zwischen den Basislerner senkt, erhöht es die Varianz σ^2 jedes einzelnen von ihnen. Die Anzahl der jedem Basislerner angebotenen Merkmale ist daher abstimmenswert: zu viele, und die Entscheidungsbäume stimmen überein; zu wenige, und jeder von ihnen sagt schlecht voraus.

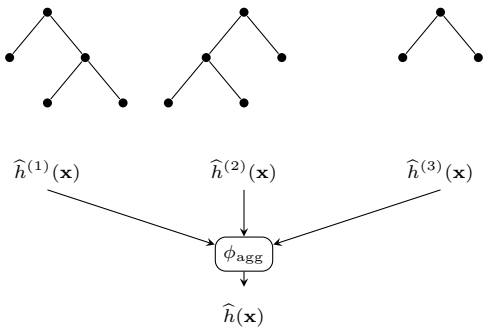


Abbildung 2: Ein Random Forest mit drei Basislerner, jeder lernt einen Entscheidungsbaum auf seiner eigenen zufälligen Auswahl der Datenpunkte des Trainingsdatensatz. Für einen Merkmalsvektor $\mathbf{x}$ wird die Vorhersage $\widehat{h}(\mathbf{x})$ aus einer Aggregationsregel ϕ_{agg} erhalten, die auf die Baum-Vorhersagen $\widehat{h}^{(1)}(\mathbf{x}), \ldots, \widehat{h}^{(3)}(\mathbf{x})$ angewendet wird.

Ein Random Forest liefert auch einen Merkmalswichtigkeit-Wert für jedes Merkmal: den Anteil der gesamten Splits, die dieses Merkmal verwenden. Für die Bank benennen diese Werte die Attribute eines Antrags, die ihr Forest am häufigsten testet. Abb. 3 berechnet diese Werte für einen Forest, der auf vier Merkmale eines Tages in Krems an der Donau trainiert wurde.

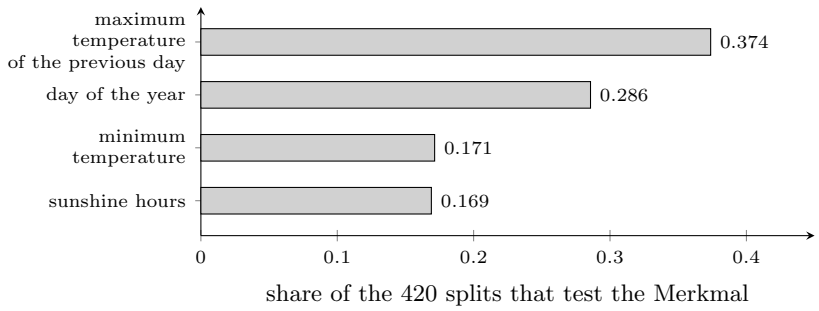


Abbildung 3: Merkmalswichtigkeit-Werte eines Forests von 60 Entscheidungsbäume, der auf vier Merkmale eines Tages in Krems an der Donau trainiert wurde, um die maximale Lufttemperatur vorherzusagen. Jedem Basislerner wurden zufällig ausgewählte Tage und ein zufällig ausgewähltes Paar der vier Merkmale angeboten, und jeder von ihm vorgenommene Split wurde aufgezeichnet. Der Wert eines Merkmal ist der Anteil der 420 Splits, die es testen. Daten generiert von `pythondemos/randomforest.py`

Siehe auch: Ensemble, Bagging, Entscheidungsbaum, Bootstrap, Merkmalswichtigkeit, Varianz, Korrelation.

Querverweise

Literatur

1. Breiman (2001). Random Forests. Machine Learning, 45(1), pp. 5–32.

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