Ensemble

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

Eine Ensemble-Methode kombiniert mehrere Methoden des maschinellen Lernens, die jeweils als Basislerner bezeichnet werden. Ensemble-Methoden unterscheiden sich darin, wie sie die Basislerner konstruieren und wie sie deren Ergebnisse kombinieren. Bagging trainiert jeden Basislerner auf einer durch Resampling erzeugten Kopie des Trainingsdatensatzes. Boosting führt sie sequenziell aus, wobei jeder an die Fehler der vorherigen angepasst wird. Stacking verwendet Basislerner, um verschiedene Modelle auf demselben Trainingsdatensatz zu trainieren, und erlernt eine Aggregationsregel. Für den quadratischen Fehlerverlust ist der Gewinn exakt: Der Fehler der gemittelten Vorhersage ist der durchschnittliche Fehler der Basislerner abzüglich ihrer durchschnittlichen Abweichung untereinander, sodass es sich gerade dann lohnt, ein Ensemble zu bilden, wenn sich die Basislerner uneinig sind.

Definition

Die Vorhersage der Lufttemperatur um 15:00 Uhr (UTC) eines Tages an der Wetterstation Krems an der Donau (Österreich) aus dem Temperaturmesswert von 06:00 Uhr desselben Tages kann mit einem linearen Modell, mit einem Entscheidungsbaum oder mit einem künstlichen neuronalen Netz erfolgen. Abb. 1 zeigt alle drei, gelernt auf 244 der 365 Tage des Jahres 2020. Keines davon ist überall am besten: Die lineare Hypothese verfehlt die Krümmung am kalten Ende, der Entscheidungsbaum antwortet mit einer Treppenfunktion, und das künstliche neuronale Netz krümmt sich dort, wo die anderen beiden es nicht können. Eine Modellauswahl würde eines davon behalten und die übrigen verwerfen. Die Kombination dieser Basislerner kann bessere Ergebnisse liefern als die Auswahl eines einzelnen.

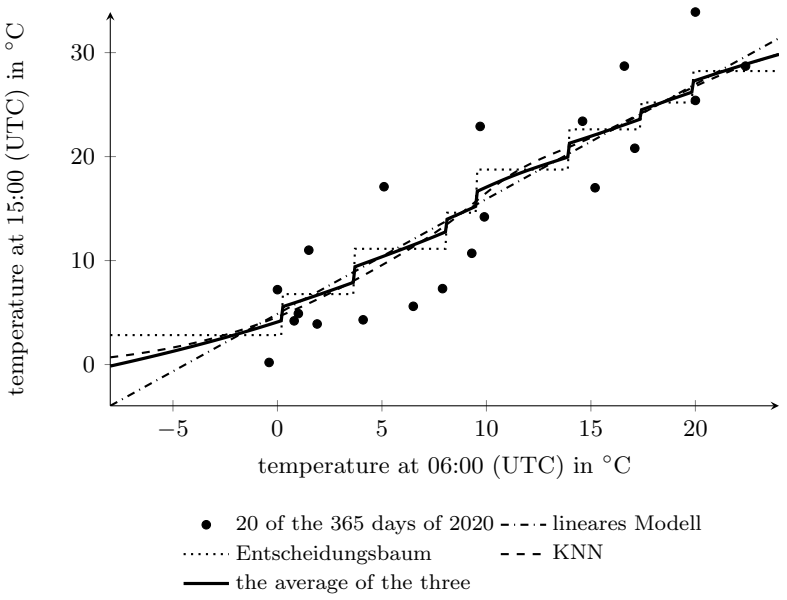


Abbildung 1: Drei Basislerner für dieselbe Aufgabe: die Lufttemperatur um 15:00 Uhr (UTC) eines Tages in Krems an der Donau aufgetragen gegen den Messwert von 06:00 Uhr desselben Tages. Die Punkte sind 20 der 365 Tage des Jahres 2020, jeder 19. Tag. Das lineare Modell (strichpunktierte Kurve), der Entscheidungsbaum (gepunktete Kurve) und das künstliche neuronale Netz (gestrichelte Kurve) werden jeweils auf denselben 244 Trainingstagen gelernt. Auf den 121 zurückgehaltenen Tagen betragen ihre mittleren quadratischen Fehler 20.53, 21.94 und 20.22, sodass keiner von ihnen allein am besten abschneidet; der Mittelwert der drei (durchgezogene Kurve) erreicht 20.46. Daten generiert durch `pythondemos/ensemble.py`

Die Kombination mehrerer Methoden des maschinellen Lernens zu einem Prädiktor ist eine Ensemble-Methode: Jede kombinierte Methode ist ein Basislerner (Hastie et al., 2009, Ch. 16). Die Basislerner können auf empirischer Risikominimierung basieren und unterschiedliche Entscheidungen für die Verlust, das Modell und den Trainingsdatensatz nutzen. Von M Basislerner lernt der Basislerner ℓ eine Hypothese $\widehat{h}^{(\ell)}$ aus seinem Hypothesenraum $\mathcal{H}^{(\ell)}$ und seinem möglicherweise modifizierten Trainingsdatensatz $\widetilde{\mathcal{D}}^{(\ell)}$, und eine Aggregationsregel ϕ_{agg} kombiniert die Ausgaben dieser Hypothesen zur endgültigen Vorhersage. Die Aggregation kann auf eine Mittelwertbildung (bei der Regression), wie in Abb. 1, oder auf einen Mehrheitsentscheid (bei der Klassifikation) hinauslaufen; Abb. 2 zeigt den allgemeinen Aufbau.

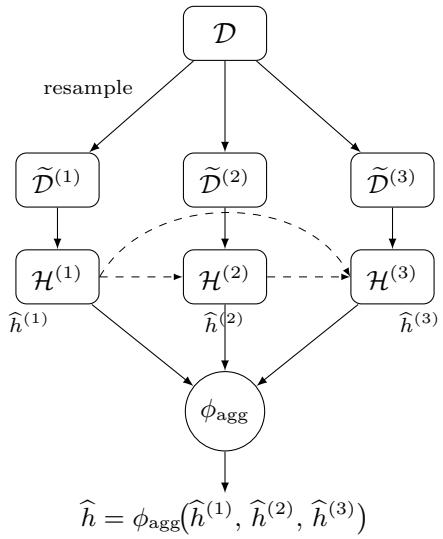


Abbildung 2: Allgemeines Ensemble mit drei Basislerner: Der Basislerner $\ell \in \{1, 2, 3\}$ nutzt ERM, um $\widehat{h}^{(\ell)} \in \mathcal{H}^{(\ell)}$ aus dem modifizierten Trainingsdatensatz $\widetilde{\mathcal{D}}^{(\ell)}$ zu lernen. Ein Basislerner kann auch die Ausgabe anderer Basislerner verwenden (gestrichelte Pfeile). Die endgültige Hypothese $\widehat{h}$ wird durch Aggregation der Hypothesen der Basislerner über eine Aggregationsregel ϕ_{agg} gewonnen

Was die Kombination gewinnt, ist genau die Abweichung zwischen den Basislerner. Für den quadratischen Fehlerverlust ist der Fehler der gemittelten Vorhersage gleich dem durchschnittlichen Fehler der Basislerner abzüglich ihrer durchschnittlichen Abweichung untereinander, wobei die Abweichung eines Basis-Lerners die quadrierte Differenz zwischen seiner eigenen Vorhersage und der gemittelten ist (Krogh and Vedelsby, 1994). Die Zerlegung ist eine Identität, keine Näherung: Auf den zurückgehaltenen Tagen aus Abb. 1 ergibt der durchschnittliche Einzelfehler von 20.90 abzüglich der durchschnittlichen Abweichung von 0.43 genau den Wert 20.46, den die gemittelte Vorhersage tatsächlich aufweist. Daraus folgen zwei Konsequenzen. Der Mittelwert ist niemals schlechter als seine Basislerner im Durchschnitt, da eine solche Abweichung nicht negativ sein kann. Und Basislerner, die überall übereinstimmen, bringen keinen Gewinn, weshalb sie mit verschiedenen Modelle oder auf verschiedenen Datenpunkte trainiert werden. Den *besten* einzelnen Basislerner zu übertreffen, ist eine stärkere Aussage, die durch die Identität nicht garantiert wird: Über 50 Wiederholungen auf zufälligen Hälften des Trainingsdatensatzes gelang dies dem Mittelwert 22-mal.

Derselbe Sachverhalt ausgedrückt durch die Varianz: Betrachtet seien M Basislerner, deren Vorhersagen identisch verteilt sind, jeweils mit der Varianz σ^2 und paarweiser Korrelation ρ. Der Mittelwert dieser Vorhersagen besitzt die

Varianz (Hastie et al., 2009, Sect. 15.2)

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

Der zweite dieser beiden Summanden verschwindet mit wachsendem M, während der erste, bestimmt durch die Korrelation zwischen den Basislerner, begrenzt, was durch Mittelwertbildung erreicht werden kann.

Ensemble-Methoden unterscheiden sich darin, wie sie die Basislerner konstruieren und wie sie deren Ergebnisse kombinieren. Bagging trainiert jeden Basislerner auf einer durch zufälliges Resampling gezogenen Stichprobe des Trainingsdatensatzes und mittelt sie; dies reduziert die Varianz eines instabilen Basis-Lerners wie etwa eines tiefen Entscheidungsbaums, während der systematische Teil seines Fehlers, die Verzerrung, unverändert bleibt (Hastie et al., 2009, Sect. 8.7). Ein Random Forest ist Bagging von Entscheidungsbäume mit einer zusätzlichen zufälligen Auswahl der an jeder Verzweigung berücksichtigten Merkmale, was die Korrelation zwischen den Bäumen senkt (Breiman, 2001, Sect. 3; Hastie et al., 2009, Sect. 15.2).

Boosting führt die Basislerner sequenziell aus, wobei jeder an die Fehler der vorherigen angepasst wird. Die Anzahl der Durchläufe ist ein Hyperparameter: Der Trainingsfehler nimmt weiter ab, während der Validierungsfehler schließlich ansteigt, sodass Boosting zur Überanpassung neigt, wenn es zu lange ausgeführt wird (Hastie et al., 2009, Sect. 10.12). Stacking trainiert die Basislerner auf demselben Trainingsdatensatz mit unterschiedlichen Modelle und erlernt die Aggregationsregel selbst mithilfe einer zweiten Methode des maschinellen Lernens. Die zweite Methode wird auf Vorhersagen für Datenpunkte trainiert, die nicht zu den Trainingsdatensätze der Basislerner gehörten. Andernfalls würden diejenigen Basislerner bevorzugt, die sich am besten an den Trainingsdatensatz anpassen, unabhängig davon, wie gut sie generalisieren (Hastie et al., 2009, Sect. 8.8).

Siehe auch: Bagging, Boosting, Stacking, Basislerner, Random Forest, Varianz, Korrelation.

Querverweise

Literatur

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

2. Krogh and Vedelsby (1994). Neural Network Ensembles, Cross Validation, and Active Learning. In: Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 231–238.

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