Modellauswahl

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

Die Modellauswahl wählt unter mehreren alternativen machine learning (ML)-Methoden, indem sie deren Validierungsfehler vergleicht. Jede ML-Methode lernt eine Hypothese über eine Training-Methode wie empirische Risikominimierung (ERM). Eine grundlegende Entwurfsentscheidung einer ML-Methode ist der zugrunde liegende Hypothesenraum oder das Modell. Der effektive Hypothesenraum einer ML-Methode kann auch von Hyperparameter der optimization-Algorithmen abhängen, die zur Lösung von ERM verwendet werden. Der Trainingsfehler einer ML-Methode kann nicht das Kriterium für die Modellauswahl sein, da er immer den größten von verschachtelten Hypothesenräume wählt. Der Validierungsfehler kann aus dem durchschnittlichen Verlust auf einem einzelnen Validierungsdatensatz oder über k-fache Kreuzvalidierung erhalten werden. Die ausgewählte Hypothese muss auf Datenpunkte bewertet werden, die weder zum Lernen einer einzelnen Hypothese noch zur Auswahl unter ihnen mittels eines Validierungsdatensatz verwendet wurden.

Definition

Abb. 1 zeigt Datenpunkte, die von einer Wetterstation in Krems an der Donau (Österreich) gesammelt wurden. Jeder Datenpunkt repräsentiert einen Tag, mit der minimum Tagestemperatur als Merkmal x und der maximum Tagestemperatur als Label y. Die Abbildung zeigt auch zwei Hypothese-Abbildungen, die jeweils von einer machine learning (ML)-Methode über empirische Risikominimierung (ERM) aus einem Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ gelernt wurden, das aus sechs Datenpunkte besteht. Eine ist eine lineare Hypothese $\widehat{h}^{(1)}$, gelernt über ein lineares Modell. Die andere ist ein Polynom fünften Grades $\widehat{h}^{(2)}$, gelernt über den größeren Hypothesenraum der polynomial regression.

Das Polynom $\widehat{h}^{(2)}$ liefert perfekte Vorhersagen auf jedem Datenpunkt von $\mathcal{D}^{(\text{train})}$, sodass sein Trainingsfehler, der durchschnittliche squared error loss über $\mathcal{D}^{(\text{train})}$, null ist. Im Gegensatz dazu führt die lineare Hypothese $\widehat{h}^{(1)}$ zu einem Trainingsfehler von 0.47. Der Trainingsfehler reiht das Polynom daher an erster Stelle, und eine Befolgung dieser Reihung würde den Kandidaten auswählen, der auf einem Datenpunkt außerhalb von $\mathcal{D}^{(\text{train})}$ am schlechtesten vorhersagt. Die Modellauswahl reiht die beiden stattdessen nach dem Validierungsfehler E_v, der auf einem Validierungsdatensatz $\mathcal{D}^{(\text{val})}$ berechnet wird,

das disjunkt vom Trainingsdatensatz $\mathcal{D}^{(\text{train})}$ ist. Auf den vierzehn validation-Datenpunkte, die in Abb. 1 gezeigt werden, beträgt der durchschnittliche squared error loss 2065 für das Polynom und 7.06 für die lineare Hypothese, sodass die Modellauswahl die lineare Hypothese $\widehat{h}^{(1)}$ wählt.

Der Validierungsfehler der ausgewählten Hypothese ist nach unten verzerrt, da die Auswahl den Kandidaten mit dem kleinsten bevorzugte. Im Kontext von Abb. 1 betrachte man zwölf neue Tage, die wie die Datenpunkte dieser Abbildung generiert wurden, wobei sechs davon zur Auswahl unter sechs Kandidatengraden und sechs zur Bewertung des ausgewählten Kandidaten verwendet wurden. Über 2000 zufällige Ziehungen dieser zwölf Tage beträgt der kleinste durchschnittliche squared error loss auf den sechs zur Auswahl verwendeten Tagen im Durchschnitt 4.70, während der durchschnittliche squared error loss des ausgewählten Kandidaten auf den sechs anderen Tagen 9.61 beträgt. Ein Testdatensatz $\mathcal{D}^{(\text{test})}$, das sowohl von $\mathcal{D}^{(\text{train})}$ als auch von $\mathcal{D}^{(\text{val})}$ disjunkt ist, wird daher vor der Modellauswahl beiseitegelegt. Sein durchschnittlicher Verlust schätzt das Risiko der ausgewählten Hypothese. Die sechs Datenpunkte des Testdatensatz in Abb. 1 werden zu diesem Zweck verwendet, und der durchschnittliche squared error loss auf ihnen ist die Schätzung, die in der Bildunterschrift angegeben wird.

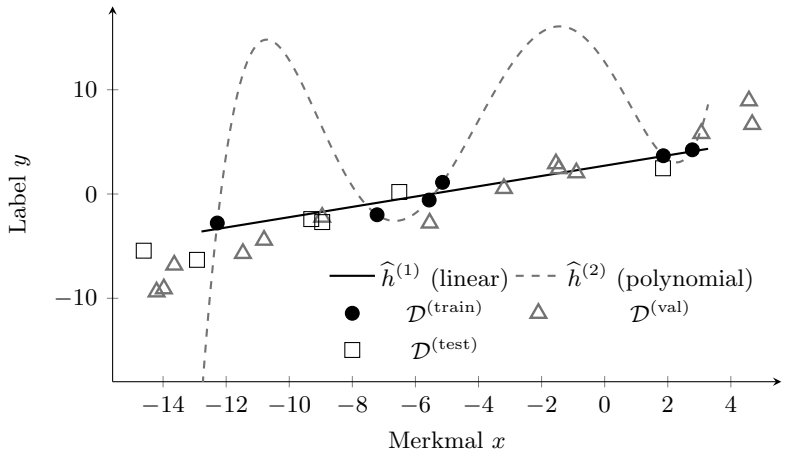


Abbildung 1: Modellauswahl zwischen zwei ML-Methoden, die jeweils eine Hypothese über ERM lernen, wobei ein unterschiedlicher Hypothesenraum verwendet wird, aber dieselben sechs Datenpunkte als Trainingsdatensatz (gefüllte Kreise). Eine Methode verwendet Polynome fünften Grades als Hypothesenraum, die andere ein lineares Modell. Das gelernte Polynom fünften Grades $\widehat{h}^{(2)}$ (gestrichelte Kurve) erreicht einen Trainingsfehler von null, während die gelernte lineare Hypothese $\widehat{h}^{(1)}$ (durchgezogene Kurve) einen Trainingsfehler von 0.47 erreicht. Die entsprechenden Validierungsfehler, berechnet als squared error loss über ein Validierungsdatensatz (offene Dreiecke), betragen 2065 für das Polynom und 7.06 für die lineare Hypothese. Basierend auf den Validierungsfehler wird die lineare Hypothese $\widehat{h}^{(1)}$ ausgewählt. Die ausgewählte Hypothese wird schließlich auf einem Testdatensatz von sechs Datenpunkte (offene Quadrate) bewertet, die weder zum Trainingsdatensatz noch zum Validierungsdatensatz gehören. Der resultierende durchschnittliche squared error loss, der von $\widehat{h}^{(1)}$ auf dem Testdatensatz verursacht wird, 1.83, schätzt das Risiko $\bar{L}\big(\widehat{h}^{(1)}\big)$. Daten generiert von `pythondemos/validation.py`

Die Modellauswahl wählt unter mehreren alternativen ML-Methoden, indem sie deren Validierungsfehler vergleicht (Shalev-Shwartz and Ben-David, 2014, Sect. 11.2.2). Betrachte C Kandidaten $\widehat{h}^{(c)}$, $c = 1, \ldots, C$, die aus ERM mit einem unterschiedlichen effektiven Hypothesenraum oder Modell $\mathcal{H}^{(c)}$ erhalten wurden. Der effektive Hypothesenraum kann auch von Hyperparameter des optimization-Algorithmus abhängen, der zur Lösung von ERM verwendet wird. Beispiele für solche Hyperparameter sind die Lernrate einer gradientenbasiertes Verfahren, die Stärke α eines Regularisierung-Strafterm (Goodfellow et al., 2016, Sect. 11.4) und die Tiefe eines Entscheidungsbaum (Hastie et al., 2009, Sect. 9.2.2).

Der Trainingsfehler der gelernten Hypothesen $\widehat{h}^{(c)}$ ist als Kriterium für die Auswahl zwischen ihnen nicht nützlich. Als Beispiel wählt er für verschachtelte Kandidaten $\mathcal{H}^{(1)} \subseteq \cdots \subseteq \mathcal{H}^{(C)}$ immer den größten Hypothesenraum: jeder Hy-

pothesenraum enthält den vorherigen, sodass der kleinste Trainingsfehler über $\mathcal{H}^{(c+1)}$ höchstens der kleinste Trainingsfehler über $\mathcal{H}^{(c)}$ ist. Dies gilt auch dann, wenn der größte Hypothesenraum zu Überanpassung führt, wie es das Polynom fünften Grades in Abb. 1 tut (Hastie et al., 2009, Sect. 7.2, Fig. 7.1; Shalev-Shwartz and Ben-David, 2014, Sect. 11.2.3). Abb. 2 zeigt dies für die polynomialen Hypothesenräume $\mathcal{H}^{(c)}$ vom Grad $c = 0, \ldots, 5$ auf dem Trainingsdatensatz und Validierungsdatensatz von Abb. 1: der Trainingsfehler fällt mit dem Grad, während der Validierungsfehler sein Minimum bei Grad eins hat.

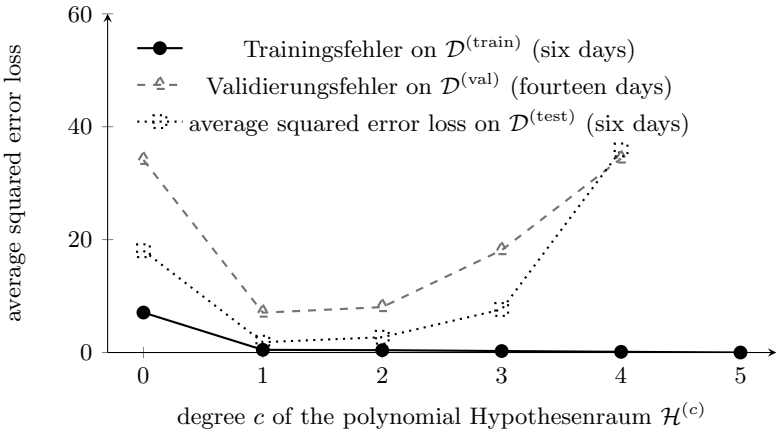


Abbildung 2: Trainingsfehler (durchgezogene Kurve, gefüllte Kreise), Validierungsfehler (gestrichelte Kurve, offene Dreiecke) und durchschnittlicher squared error loss auf dem Testdatensatz (gepunktete Kurve, offene Quadrate) der Hypothesen, die über ERM über die polynomialen Hypothesenräume $\mathcal{H}^{(c)}$ vom Grad $c = 0, \ldots, 5$ auf den Datenpunkte von Abb. 1 gelernt wurden. Der Trainingsfehler fällt mit dem Grad und erreicht null bei Grad fünf, wo das Polynom durch alle sechs Datenpunkte von $\mathcal{D}^{(\text{train})}$ verläuft. Der Validierungsfehler fällt von Grad null auf Grad eins und steigt danach an; sein Wert bei Grad fünf, 2065, liegt oberhalb des gezeichneten Bereichs, ebenso wie der durchschnittliche squared error loss auf dem Testdatensatz bei diesem Grad. Daten generiert von `pythondemos/modelsel.py`

Die Modellauswahl folgt daher dem Modell-Training, z.B. über ERM, und der Modell-validation, z.B. über k-fache Kreuzvalidierung. Algorithmus 1 fasst einen vollständigen ML-Workflow zusammen, bestehend aus Modell-Training, Modell-validation und Modellauswahl.

Algorithm 1 Modellauswahl (Shalev-Shwartz and Ben-David, 2014, Sect. 11.2.2)

Input: Trainingsdatensatz $\mathcal{D}^{(\text{train})}$, Validierungsdatensatz $\mathcal{D}^{(\text{val})}$, a Verlustfunktion L for Training and a Verlustfunktion L_v for validation, candidate Hypothesenräume $\mathcal{H}^{(c)}$ with their Hyperparameter settings, for $c = 1, \ldots, C$

1: **for** $c = 1, \ldots, C$ **do**

2: learn a Hypothese via ERM,

$$\widehat{h}^{(c)} = \arg\min_{h \in \mathcal{H}^{(c)}} \frac{1}{|\mathcal{D}^{(\text{train})}|} \sum_{\mathbf{z} \in \mathcal{D}^{(\text{train})}} L\left(\mathbf{z}, h\right)$$

3: compute the Validierungsfehler $E_v\left(\widehat{h}^{(c)}\right)$ via the average L_v of $\widehat{h}^{(c)}$ on $\mathcal{D}^{(\text{val})}$

4: **end for**

5: $c^\star := \arg\min_{c=1,\ldots,C} E_v\left(\widehat{h}^{(c)}\right)$, a tie going to the smaller Hypothesenraum

Output: the selected Hypothese $\widehat{h}^{(c^\star)}$

Die beiden in Algorithmus 1 für Training und validation verwendeten Verlustfunktionen können unterschiedlich sein: ERM benötigt eine, die seine optimization-Methode verarbeiten kann, während die validation die für die Anwendung relevante Verlustfunktion verwenden kann, wie z.B. die 0/1 loss. Wenn das Trainingsdatensatz zu klein ist, um ein Validierungsdatensatz abzuspalten, können die Schritte 2 und 3 von Algorithmus 1 durch k-fache Kreuzvalidierung ersetzt werden (Hastie et al., 2009, Sect. 7.10; Stone, 1974).

Die endgültige Hypothese $\widehat{h}^{(c^\star)}$, die von Algorithmus 1 geliefert wird, muss auf einem Testdatensatz bewertet werden. Ein Testdatensatz besteht aus Datenpunkte, die weder zum Lernen einer einzelnen $\widehat{h}^{(c)}$ noch zur Auswahl zwischen ihnen verwendet wurden. Unter einer Annahme unabhängig und identisch verteilter Daten (u.i.v.-Annahme) für die Daten-Generierung ist der durchschnittliche Verlust auf $\mathcal{D}^{(\text{test})}$ ein unbiased estimator des Risiko von $\widehat{h}^{(c^\star)}$. Die Datenpunkte von $\mathcal{D}^{(\text{test})}$ sind unabhängig davon, wie $\widehat{h}^{(c^\star)}$ erhalten wurde (Shalev-Shwartz and Ben-David, 2014, Sect. 11.2.1), sodass jeder Verlust das Risiko als seinen Erwartungswert hat, und somit auch deren Durchschnitt (Bertsekas and Tsitsiklis, 2008, Sect. 9.1). Wenn $\mathcal{D}^{(\text{train})}$ und $\mathcal{D}^{(\text{val})}$ als Zufallsvariablen (ZV) modelliert werden, so ist es auch $\widehat{h}^{(c^\star)}$, und damit das geschätzte Risiko: der Erwartungswert ist dann bedingt auf $\widehat{h}^{(c^\star)}$, nur über $\mathcal{D}^{(\text{test})}$.

Siehe auch: validation, Validierungsfehler, k-fache Kreuzvalidierung, Trainingsdatensatz, Validierungsdatensatz, Testdatensatz, Hyperparameter, Hypothesenraum, Überanpassung.

Querverweise

Literatur

1. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.

2. Goodfellow et al. (2016). Deep Learning. MIT Press, Cambridge, MA, USA.

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

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

5. Stone (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. Journal of the Royal Statistical Society. Series B (Methodological), 36(2), pp. 111–133.