Theorie des maschinellen Lernens

Hans U. Simon

26. April 2017

4 Lernen und uniforme Konvergenz (endlicher Fall)

In Abschnitt 4.1 wird gezeigt, dass die uniforme Konvergenz[1] von $L_S(h)$ gegen $L_{\mathcal{D}}(h)$ garantiert, dass die Hypothesenklasse $\mathcal{H}$ agnostisch PAC-lernbar ist. Zum Beispiel ist dann „Empirical Risk Minimization (ERM)“ eine geeignete Lernstrategie. In Abschnitt 4.2 weisen wir nach, dass alle endlichen Hypothesenklassen (bezüglich beschränkter Verlustfunktionen) agnostisch PAC-lernbar sind. In dem abschließenden Abschnitt 4.3 gehen wir kurz auf den sogenannten „bias-complexity tradeoff“ ein.

4.1 Uniforme Konvergenz als Garant für Lernbarkeit

„Empirical Risk Minimization (ERM)“ ist folgende Lernstrategie: gegeben eine Trainingsmenge $S \subseteq Z^m$, ermittle eine Hypothese $h_S \in \mathcal{H}$, die $L_S(h)$ minimiert, d.h., es gilt

$$L_S(h_S) = \min_{h \in \mathcal{H}} L_S(h) \ . \tag{1}$$

Erinnern wir uns daran, dass der Lerner eigentlich $L_{\mathcal{D}}(h)$ anstelle von $L_S(h)$ minimieren möchte (aber leider kennt er ja die Wahrscheinlichkeitsverteilung $\mathcal{D}$ nicht). Wenn $L_S(h)$ für alle $h \in \mathcal{H}$ eine gute Schätzung von $L_{\mathcal{D}}(h)$ ist, dann sollte ERM eine fast-optimale Hypothese finden, richtig? Die folgenden Ausführungen präzisieren diese Intuition.

Definition 4.1 *Eine Trainingsmenge $S \in Z^m$ heißt* ε-repräsentativ*, falls folgendes gilt:*

$$\forall h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon \ . \tag{2}$$

Lemma 4.2 *Wir setzen voraus, dass $\mathcal{H}$ endlich ist, so dass eine „optimale Hypothese“ $h^* \in \mathcal{H}$ mit $L_{\mathcal{D}}(h^*) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$ existiert. Weiter sei S $\varepsilon/2$-repräsentativ und h_S die von ERM (angewendet auf S) ermittelte Hypothese. Dann gilt $L_{\mathcal{D}}(h_S) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$.*

[1] uniform über alle Wahlen von $h \in \mathcal{H}$

Beweis Der Beweis nutzt (2) aus (mit $\varepsilon/2$ in der Rolle von ε) und die Tatsache, dass ERM den „empirischen Champion“ (im Sinne von (1)) ermittelt:

$$L_{\mathcal{D}}(h_S) \overset{(2)}{\leq} L_S(h_S) + \frac{\varepsilon}{2} \overset{(1)}{\leq} L_S(h^*) + \frac{\varepsilon}{2} \overset{(2)}{\leq} L_{\mathcal{D}}(h^*) + \varepsilon \ .$$

qed.

Definition 4.3 (uniforme Konvergenzbedingung (UKB)) *Eine Hypothesenklasse $\mathcal{H}$ erfüllt die* uniforme Konvergenzbedingung (UKB) *(bezüglich einer Verlustfunktion ℓ), falls eine Funktion $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ existiert, so dass für alle $0 < \varepsilon, \delta < 1$, alle $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ und jede Wahl von $\mathcal{D}$ folgende Bedingung erfüllt ist:*

$$\Pr_{S \sim \mathcal{D}^m}[S \text{ ist } \varepsilon\text{-repräsentativ}] \geq 1 - \delta \ .$$

Im Folgenden bezeichne $m_{\mathcal{H}}^{UC}$ die Funktion, die der Definition 4.3 genügt und unter dieser Nebenbedingung die kleinstmöglichen Werte $m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ hat.[2] Es ergibt sich unmittelbar die

Folgerung 4.4 *Wenn $\mathcal{H}$ eine endliche Hypothesenklasse ist, welche die UKB bezüglich einer Verlustfunktion ℓ erfüllt, dann ist $\mathcal{H}$ agnostisch PAC-lernbar bezüglich ℓ (und zwar mit ERM). Darüberhinaus gilt:*

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2}, \delta\right) \ .$$

Wir merken kurz an, dass für beliebige (evtl. unendliche) Klassen $\mathcal{H}$ die Folgerung 4.4 ebenso gültig ist bis auf folgende kleine Abschwächung:

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{3}, \delta\right) \ .$$

Diese Behauptung wird evtl. in den Übungen überprüft werden.

4.2 Agnostische PAC-Lernbarkeit endlicher Hypothesenklassen

Nach den Ausführungen in Abschnitt 4.1 sollte klar sein, dass wir lediglich nachweisen müssen, dass $\mathcal{H}$ die UKB erfüllt. Wir werden in diesem Abschnitt zeigen, dass dies für alle beschränkten Verlustfunktionen der Form $\ell : \mathcal{H} \times Z \to [a, b]$ der Fall ist. Durch Normierung können wir uns natürlich, wann immer uns das besser passt, auf Funktionen der Form $\ell : \mathcal{H} \times Z \to [0, 1]$ zurückziehen.
Um die UKB zu verifizieren, benötigen wir einige Hilfsmittel aus der Statistik:

[2] UC = Uniform Convergence.

„Union Bound“: Für Ereignisse $E_1, \ldots, E_s$ gilt:

$$\Pr\left[\bigcup_{i=1}^{s} E_i\right] \leq \sum_{i=1}^{s} \Pr[E_i] \ . \tag{3}$$

Im Deutschen ist diese Regel unter dem etwas sperrigen Namen „Subadditivität von Wahrscheinlichkeitsmaßen“ bekannt. Wir merken kurz an, dass sich die mengentheoretische Vereinigung, sprachlich gesehen, oft hinter einem Existenzquantor versteckt, da das Ereignis $\cup_{i=1}^{s} E_i$ eintritt genau dann, wenn ein $i \in [s]$ existiert mit: das Ereignis E_i ist eingetreten.

Hoeffding's Ungleichung: Es sei $\theta_1, \ldots, \theta_m$ eine Sequenz von unabhängigen, identisch verteilten Zufallsvariablen — im Folgenden kurz iid-Variablen genannt[3] — mit Werten in einem beschränkten Intervall $[a, b]$ $(a < b \in \mathbb{R})$ und mit Erwartungswert μ. Dann gilt für alle $\varepsilon > 0$:

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} \theta_i > \mu + \varepsilon\right] \leq \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right) \ . \tag{4}$$

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} \theta_i < \mu - \varepsilon\right] \leq \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right) \ . \tag{5}$$

$$\Pr\left[\left|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu\right| > \varepsilon\right] \leq 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right) \ . \tag{6}$$

Die Hoeffding-Ungleichungen begrenzen die Wahrscheinlichkeit, dass der empirische Mittelwert der iid-Variablen weit vom Erwartungswert abweicht. Die Funktion $\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$ konvergiert mit wachsendem m in negativ exponentieller Geschwindigkeit gegen Null. Freilich folgt (6) direkt aus (4) und (5) vermöge der „Union Bound“.

Wir wählen eine Hypothese $h \in \mathcal{H}$ beliebig aber fest aus. Mit einer zufälligen Trainingsmenge $S = (z_1, \ldots, z_m) \sim \mathcal{D}^m$ assoziieren wir die iid-Variablen $\theta_i = \ell(h, z_i)$ für $i = 1, \ldots, m$ und setzen

$$\mu = \mathbb{E}[\theta_i] = \mathbb{E}_{z\sim\mathcal{D}}[\ell(h, z_i)] = L_{\mathcal{D}}(h)$$

Beachte, dass dann

$$\frac{1}{m}\sum_{i=1}^{m} \theta_i = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i) = L_S(h) \ .$$

Da wir anfangs des Abschnittes vorausgesetzt hatten, dass die Verlustfunktion ℓ Werte im Intervall $[a, b]$ annimmt, ist dies auch der Wertebereich der Variablen θ_i. Anwendung von (6) liefert:

$$\Pr_{S\sim\mathcal{D}^m}[|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon] \leq 2\exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right) \ . \tag{7}$$

[3] iid = independent and identically distributed

Damit S ε-repräsentativ ist, muss aber das „schlechte Ereignis“ $|L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon$ nicht nur für eine fest ausgewählte Hypothese h vermieden werden, sondern wir müssen das Ereignis

$$\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon$$

vermeiden. Durch Kombination von (7) mit der „Union Bound“ erhalten wir

$$\Pr_{S \sim \mathcal{D}^m}[\exists h \in \mathcal{H} : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon] \leq 2|\mathcal{H}| \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right) \ .$$

Wenn wir den Ausdruck auf der rechten Seite dieser Ungleichung durch δ nach oben begrenzen können, haben wir im Umkehrschluss erreicht, dass S mit einer Wahrscheinlichkeit mindestens $1 - \delta$ ε-repräsentativ ist. Wir wünschen uns daher, dass

$$2|\mathcal{H}| \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right) \leq \delta \ . \tag{8}$$

Wenn wir m hinreichend groß wählen, können wir uns diesen Wunsch selber erfüllen. Durch Auflösen von (8) nach m sehen wir, dass

$$m \geq \frac{(b-a)^2 \ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2}$$

eine hinreichend große Wahl ist. Um m nicht größer als nötig zu machen, setzen wir

$$m = \left\lceil \frac{(b-a)^2 \ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil \ .$$

Die gesamte Diskussion lässt sich zusammenfassen wie folgt:

Theorem 4.5 *Eine endliche Hypothesenklasse $\mathcal{H}$ erfüllt bezüglich einer Verlustfunktion mit Wertebereich $[a, b]$ die UKB und ist daher bezüglich ℓ agnostisch PAC-lernbar (und zwar mit ERM). Darüberhinaus gilt*

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \left\lceil \frac{(b-a)^2 \ln(2|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil$$

und daher auch

$$m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2}, \delta\right) \leq \left\lceil \frac{2(b-a)^2 \ln(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil \ .$$

Für den Spezialfall $[a, b] = [0, 1]$ ergibt sich

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq \left\lceil \frac{\ln(|\mathcal{H}|/\delta)}{2\varepsilon^2} \right\rceil \quad \textit{und} \quad m_{\mathcal{H}}(\varepsilon, \delta) \leq \left\lceil \frac{2\ln(2|\mathcal{H}|/\delta)}{\varepsilon^2} \right\rceil \ .$$

4.3 Verlustwertzerlegung und „bias-complexity-tradeoff“

Der Verlustwert (= Fehlerrate im Falle von Klassifikationsproblemen) einer ERM-Hypothese h_S lässt sich zerlegen wie folgt:

$$L_{\mathcal{D}}(h_S) = \underbrace{(\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h))}_{=:\varepsilon_{app}} + \underbrace{(L_D(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h))}_{=:\varepsilon_{est}} \quad .$$

Der Term ε_{app} wird *Approximationsfehler* genannt. Er hängt nur von der Hypothesenklasse, nicht aber von der ERM-Hypothese oder der Größe der Trainingsmenge ab. Der Term ε_{est} heißt *Schätzfehler*. Er ist abhängig von der Größe der Trainingsmenge und wächst mit der Informationskomplexität von $\mathcal{H}$. Wenn wir viel Vorwissen in die Wahl von $\mathcal{H}$ legen (strong bias), also $\mathcal{H}$ sehr speziell wählen, dann wird i.A. der Approximationsfehler groß und der Schätzfehler klein sein (low information complexity). Die Hypothesen können sich dann nicht gut an die Daten anpassen — ein Phänomen, was im Englischen „underfitting“ genannt wird. Wählen wir $\mathcal{H}$ hingegen sehr ausdrucksmächtig (weak bias), dann verhält es sich genau umgekehrt (high information complexity). Die Hypothese h_S ist dann in der Gefahr, an die Daten übermäßig angepasst zu sein (etwa „Rauschen“ oder andere Zufälligkeiten in den Daten abzubilden) — ein Phänomen, was im Englischen „overfitting“ genannt wird. Wir können also den Approximationsfehler auf Kosten eines höheren Schätzfehlers verkleinern, und umgekehrt. Im Englischen spricht man von dem „bias-complexity tradeoff“. Die Kunst besteht natürlich darin, eine geeignete Balance zu finden, so dass möglichst weder „under-“ noch „overfitting“ auftritt (good fit with the data).