Theorie des maschinellen Lernens

Hans U. Simon

8. Mai 2017

7 Nichtuniforme Lernbarkeit

Im agnostischen PAC-Lernmodell muss der Lerner imstande sein, eine Hypothese auszugeben, deren Fehlerrate mit hoher Wahrscheinlichkeit nicht wesentlich größer ist, als die beste Fehlerrate, die sich mit einer Vergleichshypothese aus der vorgegebenen Hypothesenklasse erzielen lässt. Wie wir wissen sind Klassen mit unendlicher VC-Dimension in diesem Modell (nicht einmal unter der Realisierbarkeitsannahme) lernbar. Wenn wir jedoch erlauben, die zum Lernen erforderliche Größe der Trainingsmenge nicht nur von ε, δ sondern auch von der Vergleichshypothese h abhängig zu machen, dann ist eine unendliche VC-Dimension i.A. keine unüberwindliche Barriere mehr. Ansätze dieser Art führen zum sogenannten nichtuniformen Lernmodell, dass wir in Abschnitt 7.1 eingehend besprechen. Es wird sich zeigen, dass ERM im nichtuniformen Lernmodell keine brauchbare Lernstrategie ist. Die in diesem Modell erfolgreiche Strategie trägt den Namen strukturierte Risikominimierung, oder auch kurz SRM. Eine spezielle Ausprägung von SRM namens „Minimum Description Length (MDL)“ besprechen wir in Abschnitt 7.2. Schließlich kommen wir in Abschnitt 7.3 wieder auf das „bias-complexity dilemma“ zu sprechen und vergleichen diesbezüglich das uniforme mit dem nichtuniformen Lernmodell.

7.1 Das Modell der nichtuniformen Lernbarkeit

Definition 7.1 *Eine binäre Hypothesenklasse $\mathcal{H}$ heißt* nichtuniform lernbar *wenn eine Funktion $m_{\mathcal{H}} : (0,1)^2 \times \mathcal{H} \to \mathbb{R}_+$ und ein Lernalgorithmus A existieren, so dass für alle $0 < \varepsilon, \delta < 1$ sowie für jede Verteilung $\mathcal{D}$ auf $\mathcal{X} \times \{0,1\}$ folgende Bedingung erfüllt ist: A, gestartet auf einer zufälligen Trainingsmenge $S \sim D^m$ mit $m \geq 1$, ermittelt eine Hypothese $h_S \in \mathcal{H}$ und es gilt*

$$\Pr_{S \sim D^m}\left[L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}: m_{\mathcal{H}}(\varepsilon,\delta,h) \leq m} L_{\mathcal{D}}(h) + \varepsilon\right] \geq 1 - \delta \ . \tag{1}$$

Ähnlich wie beim agnostischen Lernen kann der Lerner A prinzipiell (im Sinne der Bedingung (1)) mit jeder Hypothese aus $\mathcal{H}$ gut konkurrieren, aber im Unterschied zum herkömmlichen Modell des agnostischen Lernens hängt die Größe der dazu erforderlichen Trainingsmenge nicht nur von ε und δ sondern auch von der Vergleichshypothese h ab. Das Hauptresultat in diesem Kapitel ist das folgende:

Theorem 7.2 *Folgende Aussagen zu einer binären Hypothesenklasse $\mathcal{H}$ sind äquivalent:*

1. *$\mathcal{H}$ ist nichtuniform lernbar.*
2. *$\mathcal{H}$ ist eine abzählbare Vereinigung von Klassen endlicher VC-Dimension.*
3. *$\mathcal{H}$ ist eine abzählbare Vereinigung agnostisch PAC-lernbarer Klassen.*
4. *$\mathcal{H}$ ist eine abzählbare Vereinigung von Klassen, welche die uniforme Konvergenzbedingung erfüllen.*

Beweis Wir beweisen die Äquivalenzen durch einen Ringschluss.
1.⇒2. Es sei A der Algorithmus und $m_{\mathcal{H}} : (0,1)^2 \times \mathcal{H} \to \mathbb{R}_+$ die Funktion aus Definition 7.1. Für alle $n \geq 1$ setze

$$\mathcal{H}_n = \left\{ h \in \mathcal{H} |\ m_{\mathcal{H}}\left(\frac{1}{8}, \frac{1}{5}, h\right) \leq n \right\} \ .$$

Dann hat $S \sim \mathcal{D}^n$ mit einer Wahrscheinlichkeit von mindestens $4/5$ die Eigenschaft, dass die von A ermittelte Hypothese h_S die Bedingung $L_{\mathcal{D}}(h_S) \leq \inf_{h \in \mathcal{H}_n} L_{\mathcal{D}}(h) + 1/8$ erfüllt. In anderen Worten: $\mathcal{H}_n$ ist agnostisch PAC-lernbar in dem schwächeren Sinn, dass ε bzw. δ konstant auf $1/8$ bzw. $1/5$ gesetzt werden. Wir wissen aus Folgerung 6.8 dass dies nur möglich ist, wenn $\mathcal{H}_n$ eine endliche VC-Dimension hat. Hieraus ergibt sich 2.
2.⇒3. und 3.⇒4. Diese Implikationen ergeben sich direkt aus dem Fundamentaltheorem der PAC-Lerntheorie.
4.⇒1. Es sei $\mathcal{H} = \cup_{n \geq 1} \mathcal{H}_n$ die Darstellung von $\mathcal{H}$ als Vereinigung von Klassen, welche allesamt die uniforme Konvergenzbedingung erfüllen. Es sei $m_n^{UC} : (0,1)^2 \to \mathbb{R}_+$ eine Funktion, welche die uniforme Konvergenz bezeugt. Mit anderen Worten: für alle $0 < \varepsilon, \delta < 1$ und alle $m \geq m_n^{UC}(\varepsilon, \delta)$ gilt:

$$\Pr_{S \sim \mathcal{D}^m}[\forall h \in \mathcal{H}_n : |L_S(h) - \mathcal{D}(h)| \leq \varepsilon] \geq 1 - \delta \ .$$

Wir müssen einen Algorithmus A und eine Funktion $m_{\mathcal{H}}$ mit den in Definition 7.1 genannten Eigenschaften angeben. Die Strategie ERM ist hier nicht erfolgreich anwendbar, da sie auf der uniformen Konvergenzbedingung für die gesamte Hypothesenklasse $\mathcal{H}$ basiert. Wir werden eine neue Lernstrategie einsetzen, die unter dem Namen *Strukturelle Risikominimierung (SRM)* bekannt ist. Sie benutzt die folgende Funktion:

$$\varepsilon_n^{UC}(m, \delta) = \min\{\varepsilon \in (0,1) |\ m_n^{UC}(\varepsilon, \delta) \leq m\} \ . \tag{2}$$

Diese Funktion gibt für alle $h \in \mathcal{H}_n$ gewissermaßen die beste obere Schranke für den Schätzfehler $|L_S(h) - L_{\mathcal{D}}(h)|$ an, die für eine Trainingsmenge $S \sim \mathcal{D}^m$ mit Hilfe der Funktion m_n^{UC} (mit einer Wahrscheinlichkeit von mindestens $1 - \delta$) garantiert werden kann. Aus der Definition von ε_n^{UC} kann man folgendes ablesen:

$$m \geq m_n^{UC}(\varepsilon, \delta) \Rightarrow \varepsilon_n^{UC}(m, \delta) \leq \varepsilon \ .$$

Für alle $h \in \mathcal{H}$ sein $n(h)$ minimal mit der Eigenschaft $h \in \mathcal{H}_n$. Die Grundidee von SRM ist, eine Hypothese $h \in \mathcal{H}$ nicht nur mit dem Strafterm $L_S(h)$ zu belasten sondern auch mit ihrem mutmaßlichen Schätzfehler $\varepsilon_{n(h)}(m, \delta_n)$ bezüglich eines Konfidenzparameters δ_n. Der Parameter δ_n repräsentiert die Wahrscheinlichkeit, dass $S \sim \mathcal{D}^m$ nicht $\varepsilon_{n(h)}(m, \delta_n)$-repräsentativ für $\mathcal{H}_n$ ist. Die Parameter $\delta_1, \delta_2 \ldots$ sollen sich nur zu δ akkumulieren. Zu diesem Zweck kann eine beliebige Folge $(w_n)_{n\geq 1}$ mit $w_n > 0$ und $\sum_{n\geq 1} w_n = 1$ verwendet und $\delta_n = w_n \delta$ gesetzt werden.[1] Diese Überlegungen führen zu folgendem Algorithmus:

SRM: Gegeben eine Trainingsmenge $S \in Z^m$ und ein Konfidenzparameter $0 < \delta < 1$, finde eine Hypothese $h_S \in \mathcal{H}$, welche die Kostenfunktion $L_S(h) + \varepsilon^{UC}_{n(h)}(m, w_{n(h)}\delta)$ minimiert.

Die Funktion $m_{\mathcal{H}}(\varepsilon, \delta, h)$ muss nun noch passend gewählt werden. Wir setzen

$$m_{\mathcal{H}}(\varepsilon, \delta, h) = m^{UC}_{n(h)}\left(\frac{\varepsilon}{2}, w_{n(h)}\delta\right) \quad .$$

Diese Wahl von $m_{\mathcal{H}}$ stellt sicher, dass mit einer Wahrscheinlichkeit von mindestens $1-\delta$ eine Menge $S \sim \mathcal{D}^m$ generiert wird, so dass folgende Bedingungen gelten:

1. Sie begrenzt den Schätzfehler jeder Hypothese $h \in \mathcal{H}$ durch $\varepsilon^{UC}_{n(h)}(m, w_{n(h)}\delta)$.
2. Für alle Hypothesen $h \in \mathcal{H}$ mit $m_{\mathcal{H}}(\varepsilon, \delta, h) \leq m$ begrenzt sie den Schätzfehler durch $\varepsilon/2$.

Wir setzen im Folgenden voraus, dass diese beiden Bedingungen erfüllt sind. Es genügt zu zeigen, dass hieraus

$$L_{\mathcal{D}}(h_S) \leq \inf_{h\in\mathcal{H}: m_{\mathcal{H}}(\varepsilon,\delta,h)\leq m} L_{\mathcal{D}}(h) + \varepsilon$$

abgeleitet werden kann. Sei also nun $h \in \mathcal{H}$ und $m \geq m_{\mathcal{H}}(\varepsilon, \delta, h)$. Dann folgt:

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \varepsilon^{UC}_{n(h_S)}(m, w_{n(h_S)}\delta) \leq L_S(h) + \varepsilon^{UC}_{n(h)}(m, w_{n(h)}\delta) \leq L_{\mathcal{D}}(h) + \varepsilon \quad .$$

Die erste Ungleichung gilt, da der Schätzfehler für h_S durch $\varepsilon^{UC}_{n(h_S)}(m, w_{n(h_S)}\delta)$ begrenzt ist. Die zweite Ungleichung gilt, da h_S die Kostenfunktion $L_S(h) + \varepsilon^{UC}_{n(h)}(m, w_{n(h)}\delta)$ minimiert. Die dritte Ungleichung gilt, da wegen $m \geq m_{\mathcal{H}}(\varepsilon, \delta, h)$ sowohl $\varepsilon^{UC}_{n(h)}(m, w_{n(h)}\delta)$ als auch der Schätzfehler von h durch $\varepsilon/2$ begrenzt sind. Damit ist der Beweis für die nichtuniforme Lernbarkeit von $\mathcal{H}$ abgeschlossen. **qed.**

Mögliche Wahlen für die Folge $(w_n)_{n\geq 1}$ sind zum Beispiel $w_n = 2^{-n+1}$ oder $w_n = \frac{1}{n(n+1)}$. Diesen Wahlen von w_n liegen die geometrische und die streng harmonische Reihe mit

$$\sum_{n\geq 1} 2^{-n} = 1 = \sum_{n\geq 1} \frac{1}{n(n+1)}$$

zugrunde.

Es gibt (sogar viele) Hypothesenklassen, die nicht (uniform) PAC-lernbar sind (d.h. eine unendliche VC-Dimension besitzen) aber ohne Weiteres nichtuniform gelernt werden können:

[1] ein Trick, der unter der Bezeichnung „Aufteilung des Konfidenzparameters δ auf unendlich viele schlechte Ereignisse“ bekannt ist

Beispiel 7.3 *Ein Polynom p in einer reellen Variable zerlegt $\mathbb{R}$ in den Bereich $\{x|\ p(x) \geq 0\}$ und $\{x|\ p(x) < 0\}$. Es ist nicht schwer zu zeigen, dass die VC-Dimension der Klasse*

$$\mathcal{H}_d = \{\mathbb{1}_{[p(x)\geq 0]}|\ p \text{ ist Polynom vom Grad } d\}$$

gleich $d+1$ ist, so dass $\mathcal{H} = \cup_{d\geq 1}\mathcal{H}_d$ eine unendliche VC-Dimension besitzt. Nach dem Fundamentalsatz der PAC-Lerntheorie ist $\mathcal{H}$ nicht PAC-lernbar. Hingegen folgt aus Theorem 7.2, dass $\mathcal{H}$ nichtuniform lernbar ist.

7.2 Minimum Description Length

Es sei $\mathcal{H} = \{h_n|\ n \geq 1\}$ eine abzählbar unendliche Klasse von binären Hypothesen. Da $\mathcal{H}_n = \{h_n\}$ trivialerweise die uniforme Konvergenzbedingung erfüllt (bezeugt durch die Hoeffding-Ungleichung), ist $\mathcal{H} = \cup_{n\geq 1}\{h_n\}$ eine Darstellung von $\mathcal{H}$ als Vereinigung von Klassen, welche die uniforme Konvergenzbedingung erfüllen. Gemäß Theorem 7.2 ist $\mathcal{H}$ nichtuniform lernbar. Es hat sich somit ergeben:

Bemerkung 7.4 *Jede abzählbare (endliche oder unendliche) binäre Hypothesenklasse ist nichtuniform lernbar.*

Da es abzählbar unendliche Klassen mit unendlicher VC-Dimension gibt (zum Beispiel die Klasse aller endlichen Teilmengen der natürlichen Zahlen), gibt es abzählbar unendliche Klassen, die im nichtuniformen Modell lernbar sind, aber nicht im herkömmlichen Modell des PAC-Lernens.

Wir wollen im Folgenden untersuchen, wie SRM, angesetzt auf eine abzählbar unendliche Klasse, zu Werke geht. Zunächst einmal liefert die Hoeffding-Ungleichung für jede einzelne Hypothese $h \in \mathcal{H}$:

$$\Pr_{S\sim\mathcal{D}^m}|L_S(h) - L_{\mathcal{D}}(h) \geq \varepsilon] \leq 2e^{-2m\varepsilon^2}\ . \tag{3}$$

Auflösen der Ungleichung

$$2e^{-2m\varepsilon^2} \leq \delta$$

nach m bzw. nach ε liefert folgende mögliche Wahlen für die Funktionen m_n^{UC} bzw. ε_n^{UC}:

$$m_n^{UC}(\varepsilon,\delta) = \frac{\ln(2/\delta)}{2\varepsilon^2} \text{ und } \varepsilon_n^{UC}(m,\delta) = \sqrt{\frac{\ln(2/\delta)}{2m}}\ .$$

Wir schreiben im Folgenden $w(h_n)$ anstelle von $w(n)$ und betrachten somit w als eine auf den Hypothesen $h \in \mathcal{H}$ definierte Gewichtsfunktion. Die von SRM minimierte Kostenfunktion mit den Termen $L_S(h)$ und $\varepsilon_{n(h)}^{UC}(m, w(h)\delta)$ hat dann die Form

$$L_S(h) + \sqrt{\frac{-\ln(w(h)) + \ln(2/\delta)}{2m}}\ . \tag{4}$$

Da $\mathcal{H}$ abzählbar unendlich ist, können die Hypothesen $h \in \mathcal{H}$ durch Strings über einem Alphabet beschrieben werden. Die Beschreibung könnte zum Beispiel eine von $x \in \mathcal{X}$

abhängige mathematische Formel sein, die als Ergebnis 0 oder 1 liefert. Da letztendlich alles binär kodiert werden kann, gehen wir der Einfachheit halber davon aus, dass wir h mit einem Binärstring identifizieren können. Die Bitlänge dieses Strings notieren wir als $|h|$. Wir wollen zudem annehmen, dass $\mathcal{H}$ präfixfrei ist, d.h., keine Hypothese in $\mathcal{H}$ ist Präfix (=Anfangswort) einer anderen Hypothese in $\mathcal{H}$. Eine natürliche Wahl der Gewichte wäre nun $w(h) = 2^{-|h|}$. Diese Gewichtswahl ist zulässig, d.h., sie erfüllt die Ungleichung $\sum_{h\in\mathcal{H}} w(h) \leq 1$, wie folgendes Resultat aufzeigt:

Lemma 7.5 (Kraft's Ungleichung) *Es sei $\mathcal{H} \subseteq \{0,1\}^*$ eine präfixfreie Menge von Strings. Dann gilt $\sum_{h\in\mathcal{H}} 2^{-|h|} \leq 1$.*

Beweis Wir verbinden mit den Strings aus $\mathcal{H}$ auf folgende Weise einen Wahrscheinlichkeitsraum. Stelle einen Bitstring durch wiederholtes Werfen einer perfekten Münze her bis — es folgt die Stoppbedingung[2] — der aktuelle Bitstring ein Mitglied der Menge $\mathcal{H}$ ist. Wegen der Präfixfreiheit von $\mathcal{H}$ hat jeder String $h \in \mathcal{H}$ eine Wahrscheinlichkeit von $2^{-|h|}$ auf diese Weise zufällig generiert zu werden. Da die Generierung verschiedener Strings aus $\mathcal{H}$ disjunkte Ereignisse sind, können sich die Wahrscheinlichkeiten aller $h \in \mathcal{H}$ maximal zu 1 aufaddieren. **qed.**

Durch Einsetzen von $w(h) = 2^{-|h|}$ in (4) kann die von SRM minimierte Kostenfunktion in der Form

$$L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2m}} \tag{5}$$

geschrieben werden. Durch diese Wahl der Kostenfunktion werden im Zweifelsfalle (also bei ähnlicher empirischer Fehlerrate) Hypothesen mit einer kleinen Beschreibungslänge vor denen mit einer großen Beschreibungslänge vorgezogen. Die aus der Gewichtswahl $w(h) = 2^{-|h|}$ resultierende Variante von SRM heißt daher *MDL-Strategie* oder kurz *MDL*. „MDL" steht für „Minimum Description Length".

Wir fassen die bisherigen Überlegungen dieses Abschnittes wie folgt zusammen:
Es sei $\mathcal{H}$ eine abzählbare Hypothesenklasse, deren Hypothesen wir mit binären Beschreibungsstrings identifizieren. Somit gilt $\mathcal{H} \subseteq \{0,1\}^*$. Es wird vorausgesetzt, dass $\mathcal{H}$ präfixfrei ist. Dann führt SRM, parametrisiert mit den Gewichten $w(h) = 2^{-|h|}$, zu folgender Lernstrategie:

MDL: Gegeben eine Trainingsmenge $S \sim \mathcal{D}^m$ und ein Konfidenzparameter δ, ermittle eine Hypothese h_S, welche die Kostenfunktion (5) minimiert.

Ockham's Rasiermesser (Occam's Razor). Wilhelm von Ockham (1288–1347) war ein mittelalterlicher Philosoph in der Epoche der Spätscholastik. Er formulierte eine Art Sparsamkeitsprinzip, welches sinngemäß besagt, dass, auf der Suche nach Erklärungen für die Phänomene dieser Welt, grundsätzlich einfachen Erklärungen der Vorzug über unnötig

[2] die nicht notwendig eintreten muss

komplizierten zu geben ist. Es scheint so, als würde die MDL-Strategie exakt nach diesem Prinzip vorgehen. Dabei ist allerdings Vorsicht geboten, da wir bei der Wahl des binären Codes für die Hypothesen erst einmal völlige Narrenfreiheit besitzen. Die Tatsache, dass wir eine Hypothese h kürzer kodieren als eine andere Hypothese h' kann eine völlig willkürliche (und auch revidierbare) Festlegung sein, die nichts mit einer der Hypothese innewohnenden Einfachheit oder Komplexität zu tun haben muss. Wer an einer vertieften Diskussion solcher und verwandter Fragen interessiert ist, sollte sich die Abschnitte 7.3.1 und 7.5 des von Shalev-Shwartz und Ben-David verfassten Lehrbuches durchlesen.

7.3 Hypothesentest, uniformes und nichtuniformes Lernen

Vor dem Hintergrund von SRM als Alternative zu ERM kommen wir noch einmal auf das „bias-complexity dilemma“ zu sprechen.

Hypothesentest: Ein extrem starkes „Bias“ wäre es, sich a-priori (also vor Inspektion der empirischen Daten) auf eine einzige Hypothese h festzulegen. Die empirischen Daten dienen dann gleichsam nur als Testmenge, um a-posteriori die Fehlerrate von h empirisch zu schätzen. Der Schätzfehler $|L_S(h) - \mathcal{D}(h)|$ ist dann sehr klein und kann mit der Hoeffding Ungleichung kontrolliert werden. Auf der anderen Seite riskieren wir mit der frühen Festlegung auf eine Hypothese einen extrem hohen Approximationsfehler, da unsere Wahl der Hypothese nicht mit den empirischen Daten abgestimmt wird.

Uniformes Lernen: Ein weniger starkes „Bias“ liegt vor, wenn wir uns a-priori auf eine Hypothesenklasse (anstelle einer einzelnen Hypothese) festlegen. Wenn diese Klasse eine unendliche VC-Dimension besitzt, dann ist das „Bias“ zu schwach, und der Schätzfehler kann nicht mehr kontrolliert werden (wohingegen der Approximationsfehler vermutlich klein ist). Eine Hypothesenklasse mit endlicher VC-Dimension erfüllt, wie wir inzwischen wissen, die uniforme Konvergenzbedingung, und darüber lässt sich der Schätzfehler kontrollieren. Die Klasse muss allerdings clever gewählt sein, damit der Approximationsfehler beherrschbar ist.

Nichtuniformes Lernen: Nichtuniformes Lernen eröffnet die Möglichkeit, eine sehr ausdrucksstarke Hypothesenklasse $\mathcal{H}$ mit unendlicher VC-Dimension zu wählen, welche als eine Vereinigung $\cup_{n\geq 1}\mathcal{H}_n$ von Klassen endlicher VC-Dimension geschrieben werden kann. Um den Schätzfehler zu kontrollieren wird das von $\mathcal{H}$ repräsentierte (eher schwache) „Bias“ dadurch erhöht, dass wir a-priori eine Gewichtung der Klassen $\mathcal{H}_n$ (und damit auch der einzelnen Hypothesen) vornehmen. Wie wir im Beweis von Theorem 7.2 gesehen haben, lässt sich dadurch der Schätzfehler für alle Hypothesen aus $\mathcal{H}$ hinreichend gut kontrollieren. In das „bias-complexity dilemma“ geraten wir nun bei der Wahl der Gewichte. Wenn wir zum Beispiel die gesamte Gewichtsmasse 1 im Wesentlichen auf eine einzelne Hypothese konzentrieren, dann degeneriert das nichtuniforme Lernen zum Hypothesentesten. Nur eine clevere Wahl der Gewichte wird sowohl den Approximations- als auch den Schätzfehler unter Kontrolle halten.

Der Begriff der „cleveren Wahl“ von $\mathcal{H}$ bzw. $(w(n))_{n\geq 1}$ ist anwendungsabhängig. Eine clevere Wahl erfordert also i.A. eine große Vertrautheit mit der Anwendung, in welcher ERM bzw. SRM zum Einsatz kommt.