Theorie des maschinellen Lernens

Hans U. Simon

5. Mai 2017

6 Lernen via uniforme Konvergenz (allgemeiner Fall)

Wir diskutieren in diesem Kapitel die PAC-Lernbarkeit beliebiger (evtl. unendlich großer) binärer Hypothesenklassen und stellen die Frage, ob sie PAC-lernbar sind. Dass unendlich große Klassen grundsätzlich PAC-lernbar sein können, wird in Abschnitt 6.1 anhand eines einfachen Beispiels illustriert. In Abschnitt 6.2 beginnen wir mit allgemeineren Überlegungen und definieren zu diesem Zweck einen Parameter, genannt die VC-Dimension von $\mathcal{H}$ und notiert als $\mathrm{VCdim}(\mathcal{H})$. Wir geben zudem ein paar Beispiele zur Berechnung dieses Parameters für konkrete Hypothesenklassen an. In Abschnitt 6.3 weisen wir u.a. nach, dass $m_{\mathcal{H}}(\varepsilon, \delta)$ mindestens linear mit $\mathrm{VCdim}(\mathcal{H})$ wachsen muss. Somit sind binäre Hypothesenklassen mit unendlicher VC-Dimension nicht mal unter der Realisierbarkeitsannahme PAC-lernbar. Abschnitt 6.4 ist Sauer's Lemma gewidmet. Dieses Lemma ist ein wichtiger Baustein im Beweis, dass Klassen endlicher VC-Dimension PAC-lernbar sind. In Abschnitt 6.5 behandeln wir obere Schranken von $m_{\mathcal{H}}(\varepsilon, \delta)$, die höchstens linear mit $\mathrm{VCdim}(\mathcal{H})$ wachsen (und polynomiell in $1/\varepsilon$ und $\log(1/\delta)$). Somit sind binäre Hypothesenklassen mit endlicher VC-Dimension sogar agnostisch PAC-lernbar. Als Folgerung ergibt sich in Abschnitt 6.6 das sogenannte Fundamentaltheorem der PAC-Lerntheorie, welches die PAC-Lernbarkeit einer binären Hypothesenklasse auf mehrfache Weise charakterisiert. Der zugehörige Beweis erfordert Konzepte (wie zum Beispiel die Rademacher-Komplexität), die uns noch nicht bekannt sind. Daher werden wir diesen nur sehr grob andeuten.

6.1 Beispiel einer unendlich großen PAC-lernbaren Klasse

Die Indikatorfunktion $x \mapsto \mathbb{1}_{[x \geq a]} \in \{0, 1\}$ bildet eine reelle Zahl x genau dann auf 1 ab, falls $x \geq a$. Sie wird auch als Schwellenfunktion bezeichnet (mit Schwellwert a).

Lemma 6.1 *Die binäre Hypothesenklasse* $\mathcal{H} = \{\mathbb{1}_{[x \geq a]} |\ a \in \mathbb{R}\}$ *ist unter der Realisierbarkeitsannahme PAC-lernbar, und zwar mit ERM.*

Beweis Wegen der Realisierbarkeitsannahme dürfen wir davon ausgehen, dass ein (dem Lerner unbekanntes) $a \in \mathbb{R}$ existiert, so dass $L_{\mathcal{D}}(\mathbb{1}_{[x \geq a]}) = 0$. Es bezeichne $a_0 < a$ die größte reelle

Zahl mit $\mathcal{D}([a_0, a)) \geq \varepsilon$. Weiter bezeichne $a_1 \geq a$ die kleinste reelle Zahl mit $\mathcal{D}([a, a_1]) \geq \varepsilon$.[1] Betrachte eine Trainingsmenge $S|_{\mathcal{X}} \sim \mathcal{D}^m$. Es bezeichne $\hat{a}_0 < a$ die größte in S vorkommende mit 0 markierte Zahl (vorausgesetzt, dass S nicht nur aus positiven Beispielen besteht) und $\hat{a}_1 \geq a$ die kleinste in S vorkommende mit 1 markierte Zahl (vorausgesetzt, dass S nicht nur aus negativen Beispielen besteht). Wir betrachten die „schlechten Ereignisse“

$$B_0 \Leftrightarrow S \cap [a_0, a) = \emptyset \quad \text{und} \quad B_1 \Leftrightarrow S \cap [a, a_1] = \emptyset$$

und stellen folgende Behauptungen auf:

1. Falls weder B_0 noch B_1 eintritt, dann ist die Fehlerrate jeder ERM-Hypothese kleiner als ε.

2. Wenn m hinreichend groß gewählt wird, dann ist die Wahrscheinlichkeit für das Eintreten von $B_0 \vee B_1$ durch δ beschränkt.

Nehmen wir also erst einmal an, dass weder B_0 noch B_1 eintritt. In diesem Fall gilt $a_0 \leq \hat{a}_0 < a \leq \hat{a}_1 \leq a_1$. Eine ERM-Hypothese h_S ist auf S fehlerfrei und hat daher die Form $h_S = \mathbb{1}_{[x \geq \hat{a}]}$ für ein $\hat{a}_0 < \hat{a} \leq \hat{a}_1$. Wir betrachten im Folgenden den Fall, dass $a \leq \hat{a} \leq \hat{a}_1$. (Die Argumentation im Fall $\hat{a}_0 < \hat{a} < a$ verläuft völlig analog.) Dann ist $[a, \hat{a})$ die Fehlermenge von h_S. Wegen $[a, \hat{a}) \subseteq [a, \hat{a}_1) \subset [a, a_1]$ und der Wahl von a_1 als kleinste reelle Zahl $\mathcal{D}([a, a_1]) \geq \varepsilon$ folgt $\mathcal{D}([a, \hat{a})) \leq \varepsilon$.
Bleibt noch zu analysieren wie groß m gewählt werden muss. Die Wahrscheinlichkeiten von B_0 und B_1 sind jeweils durch $\delta/2$ beschränkt, sofern die Bedingung

$$(1 - \varepsilon)^m < e^{-\varepsilon m} \leq \frac{\delta}{2}$$

erfüllt ist. Auflösen nach m liefert die hinreichende Bedingung $m \geq \ln(2/\delta)/\varepsilon$. Es reicht also aus,

$$m = \lceil \frac{\ln(2/\delta)}{\varepsilon} \rceil$$

zu setzen. **qed.**

6.2 Definition der VC-Dimension und Beispiele

Es bezeichne wieder $\mathcal{H}$ eine binäre Hypothesenklasse über dem Grundbereich $\mathcal{X}$ und für $M \subseteq \mathcal{X}$ bezeichne $\mathcal{H}_M$ die Menge der auf $M \subseteq \mathcal{X}$ eingeschränkten Abbildungen aus $\mathcal{H}$.

Definition 6.2 (aufspaltbare Mengen) *$M \subseteq \mathcal{X}$ heißt* aufspaltbar *durch $\mathcal{H}$, wenn $\mathcal{H}_{\mathcal{M}}$ die Menge aller Abbildungen von M nach $\{0, 1\}$ ist.*

[1] Wenn a_0 bzw. a_1 nicht existiert, dann gilt $\mathcal{D}((-\infty, a)) < \varepsilon$ bzw. $\mathcal{D}([a, \infty)) < \varepsilon$, was den im Folgenden ausgeführten Beweis eher vereinfacht.

Intuitiv bedeutet das, dass jedes „Bitmuster“ auf M durch eine Abbildung in $\mathcal{H}$ realisiert werden kann.

Definition 6.3 (VC-Dimension) *Die* VC-Dimension *von $\mathcal{H}$, notiert als VCdim($\mathcal{H}$), ist definiert als die Anzahl der Elemente einer größten durch $\mathcal{H}$ aufspaltbaren Teilmenge von $\mathcal{X}$ (bzw. als ∞, falls beliebig große durch $\mathcal{H}$ aufspaltbare Teilmengen von $\mathcal{X}$ existieren). Formal:*

$$VCdim(\mathcal{H}) = \sup\{|M| : \ M \text{ ist durch } \mathcal{H} \text{ aufspaltbar}\} \ .$$

Der Nachweis von $\text{VCdim}(\mathcal{H}) = d$ kann nach folgendem Schema geschehen:

1. Wir geben eine Menge $M \subseteq \mathcal{X}$ der Größe d an, auf der sich alle 2^d Bitmuster von $\mathcal{H}$ realisieren lassen (was $\text{VCdim}(\mathcal{H}) \geq d$ beweist).

2. Wir weisen nach, dass für jede Menge $M \subseteq \mathcal{X}$, der Größe $d+1$ ein Bitmuster $b \in \{0,1\}^{d+1}$ existiert, das von $\mathcal{H}$ nicht realisiert werden kann (was $\text{VCdim}(\mathcal{H}) \leq d$ beweist).

Wir betrachten ein paar konkrete Beispielklassen. Beachte dabei, dass eine Familie von Teilmengen von $\mathcal{X}$ auch als binäre Hypothesenklasse mit Grundbereich $\mathcal{X}$ aufgefasst werden kann, indem wir $H \subseteq \mathcal{X}$ mit der Indikatorfunktion $\mathbb{1}_{[x \in H]}$ identifizieren.

Schwellenfunktionen Die Klasse der Schwellenfunktionen hat VC-Dimension 1.
Offensichtlich wird die Menge $\{0\}$ von dieser Klasse aufgespalten. Daher ist die VC-Dimension mindestens 1. Seien $x_1 < x_2$ beliebige reelle Zahlen. Dann kann das Bitmuster $(1,0)$ nicht von einer Funktion vom Typ $\mathbb{1}_{[x \geq a]}$ realisiert werden. Daher kann die VC-Dimension den Wert 1 nicht überschreiten.

Intervalle: Die Klasse der Intervalle der Form $[a,b] \subset \mathbb{R}$ hat VC-Dimension 2.
Offensichtlich wird die Menge $\{-1,1\}$ (wie jede andere 2-elementige Menge) von Intervallen aufgespalten. Daher ist die VC-Dimension mindestens 2. Seien $x_1 < x_2 < x_3$ beliebige reelle Zahlen. Dann kann das Bitmuster $(1,0,1)$ nicht von einem Intervall realisiert werden. Daher kann die VC-Dimension den Wert 2 nicht überschreiten.

Achsenparallele Rechtecke: Die Klasse der (topologisch abgeschlossenen) achsenparallelen Rechtecke in $\mathbb{R}^2$ hat VC-Dimension 4.
Offensichtlich wird die Punktmenge $\{(-1,0),(0,1),(1,0),(0,-1)\}$ durch achsenparallele Rechtecke aufgespalten. Daher ist die VC-Dimension mindestens 4. Seien $z_1, \ldots, z_5$ beliebige fünf verschiedene Punkte in $\mathbb{R}^2$. Wir wählen für jede Himmelsrichtung (Westen, Norden, Osten, Süden) jeweils einen der fünf Punkte mit einer extremalen Lage in dieser Richtung aus (zum Beispiel einen nördlichsten Punkt). Ggf. nach Umnummerierung seien z_1, z_2, z_3, z_4 die vier ausgewählten Punkte. Es sei R das kleinste achsenparallele Rechteck, das die vier ausgewählten Punkte enthält. Man überlegt sich leicht, dass z_5 in R liegen muss. Daher ist das Bitmuster $(1,1,1,1,0)$ von keinem achsenparallelen Rechteck realisierbar und die VC-Dimension kann den Wert 4 nicht überschreiten.

Konvexe Polygone: Die Klasse der konvexen Polygone mit k Ecken hat VC-Dimension $2k + 1$. Der Nachweis hierfür wird evtl. in den Übungen geführt.

Singleton-Mengen: Die Klasse $\mathcal{H}$ der 1-elementigen Teilmengen einer Grundmenge $\mathcal{X}$ mit $|\mathcal{X}| \geq 2$ hat VC-Dimension 1.
Offensichtlich wird jede einelementige Teilmenge von $\mathcal{X}$ durch $\mathcal{H}$ aufgespalten. Somit gilt $\text{VCdim}(\mathcal{H}) \geq 1$. Auf zwei beliebigen Punkten $x_1 \neq x_2 \in \mathcal{X}$ kann das Bitmuster $(1, 1)$ von $\mathcal{H}$ nicht realisiert werden. Somit gilt $\text{VCdim}(\mathcal{H}) \leq 1$.

Potenzmengen: Die Klasse $P(\mathcal{X})$ aller Teilmengen einer d-elementigen Grundmenge $\mathcal{X}$ hat VC-Dimension d.
Freilich kann jedes Bitmuster auf $\mathcal{X}$ von $P(\mathcal{X})$ realisiert werden. Daher ist die VC-Dimension mindestens d. Da $\mathcal{X}$ keine Teilmenge mit $d + 1$ Elementen hat, kann die VC-Dimension den Wert d nicht überschreiten.

Wir machen ein paar abschließende Bemerkungen:

- Es werden mindestens 2^d verschiedene Hypothesen in $\mathcal{H}$ benötigt, um alle 2^d Bitmuster über einer d-elementigen Teilmenge von $\mathcal{X}$ zu realisieren. Mit $d = \text{VCdim}(\mathcal{H})$ ergibt sich hieraus $|\mathcal{H}| \geq 2^d$ oder, anders ausgedrückt, $\text{VCdim}(\mathcal{H}) \leq \log(|\mathcal{H}|)$. Diese obere Schranke für $\text{VCdim}(\mathcal{H})$ ist gelegentlich scharf (wie zum Beipiel bei der Potenzmenge). Andererseits gibt es Klassen beliebiger (evtl. sogar unendlicher) Größe, deren VC-Dimension lediglich 1 beträgt (was wir am Beispiel der Singleton-Mengen sehen). Die Lücke zwischen $\text{VCdim}(\mathcal{H})$ und der oberen Schranke $\log(|\mathcal{H}|)$ kann also beliebig groß werden.

- Man könnte anhand vieler konkreter Klassen $\mathcal{H}$ den Eindruck gewinnen, dass die VC-Dimension von $\mathcal{H}$ übereinstimmt mit der Anzahl der Parameter, die zum Spezifizieren der Elemente von $\mathcal{H}$ benötigt werden (ein Schwellwert zur Spezifikation einer Schwellenfunktion, zwei Randpunkte eines Intervalls, vier Ecken eines achsenparallelen Rechtecks, usw.). In den Übungen werden wir sehen, dass es aber Klassen unendlicher VC-Dimension gibt, deren Elemente mit lediglich einem reellen Parameter spezifizierbar sind.

6.3 Untere Schranken zur Informationskomplexität

In diesem Abschnitt sei $\mathcal{H}$ eine binäre Hypothesenklasse der VC-Dimension d über einem Grundbereich $\mathcal{X}$. Wir betrachten PAC-Lernen von $\mathcal{H}$ unter der Realisierbarkeitsannahme. Unser Ziel ist es, die folgende untere Schranke

$$m_{\mathcal{H}}(\varepsilon, \delta) \geq \Omega\left(\frac{1}{\varepsilon}\left(d + \ln\left(\frac{1}{\delta}\right)\right)\right) \tag{1}$$

für das asymptotische Wachstum von $m_{\mathcal{H}}(\varepsilon, \delta)$ in den Parametern d, ε, δ nachzuweisen.

Untere Schranken für $m_{\mathcal{H}}(\varepsilon, \delta)$ gelten natürlich erst recht für das PAC-Lernen, wenn sie sogar für abgeschwächte Erfolgskriterien gelten, welche das Leben für den Lerner tendenziell leichter machen. Dies ist insbesondere dann von Interesse, wenn die Analyse sich unter den veränderten Modellannahmen leichter und übersichtlicher gestalten lässt. In diesem Abschnitt machen wir von dieser Technik wie folgt Gebrauch:

1. Wir fixieren Wahrscheinlichkeitsverteilungen $\mathcal{D}$ auf $\mathcal{X}$ und P auf $\mathcal{H}$.
2. Wir generieren eine zunächst unmarkierte Trainingsmenge $M = S|_{\mathcal{X}} \sim \mathcal{D}^m$.
3. Wegen der Realisierbarkeitsannahme dürfen wir uns vorstellen, dass die Labels zu M gemäß einer (dem Lerner unbekannten) Funktion $h \in \mathcal{H}$ erstellt werden. Für diesen Job benutzen wir eine zufällige Hypothese $h \sim P$. Die resultierende markierte Trainingsmenge nennen wir S.
4. Der Lerner erhält S und muss seine Hypothese h_S abliefern. Das Erfolgskriterium für h_S lautet
$$\Pr_{S|_{\mathcal{X}} \sim \mathcal{D}^m, h \sim P}[L_{\mathcal{D},h}(h_S) \leq \varepsilon] \geq 1 - \delta \ . \tag{2}$$

Wir demonstrieren nun, dass das Kriterium (2) leichter (oder zumindest nicht schwerer) zu erreichen ist, als das Erfolgskriterium

$$\forall h \in \mathcal{H} : \Pr_{S|_{\mathcal{X}} \sim \mathcal{D}^m}[L_{\mathcal{D},h}(h_S) \leq \varepsilon] \geq 1 - \delta \tag{3}$$

im PAC-Lernmodell.

Lemma 6.4 *Ein Algorithmus A, der $\mathcal{H}$ unter der Realisierbarkeitsannahme PAC-lernt erfüllt das Erfolgskritrium (2).*

Beweis Die Implikation (3)⇒(2) ergibt sich aus folgender Rechnung:

$$\begin{aligned}
\Pr_{S|_{\mathcal{X}} \sim \mathcal{D}^m, h \sim P}[L_{\mathcal{D},h}(h_S) \leq \varepsilon] &= \mathbb{E}_{S|_{\mathcal{X}} \sim \mathcal{D}^m, h \sim P}[\mathbb{1}_{[L_{\mathcal{D},h}(h_S) \leq \varepsilon]}] \\
&\geq \inf_{h \in \mathcal{H}} \mathbb{E}_{S|_{\mathcal{X}} \sim \mathcal{D}^m}[\mathbb{1}_{[L_{\mathcal{D},h}(h_S) \leq \varepsilon]}] \\
&= \inf_{h \in \mathcal{H}} \Pr_{S|_{\mathcal{X}} \sim \mathcal{D}^m}[L_{\mathcal{D},h}(h_S) \leq \varepsilon] \ .
\end{aligned}$$

qed.

Lemma 6.4 liefert die Rechtfertigung, zum Nachweis unterer Schranken für $m_{\mathcal{H}}(\varepsilon, \delta)$ mit der oben vorgestellten Variante des Lernmodells zu arbeiten.

Definition 6.5 $\mathcal{H}$ *heißt* nicht-trivial, *wenn $h_1, h_2 \in \mathcal{H}$ und $x_1, x_2 \in \mathcal{X}$ existieren, so dass*

$$h_1(x_1) = h_2(x_1) \quad \textit{und } h_1(x_2) \neq h_2(x_2) \ . \tag{4}$$

Wir haben schon des Öfteren die Ungleichung $1 + a \leq e^a$ (mit Gleichheit nur für $a = 0$) verwendet. Es ist nicht schwer zu zeigen, dass umgekehrt auch Folgendes richtig ist:

$$\forall 0 \leq a \leq \frac{1}{2} : 1 - a \geq e^{-2a} \quad . \tag{5}$$

Das folgende Resultat liefert eine erste untere Schranke für $m_{\mathcal{H}}(\varepsilon, \delta)$, die für nicht-triviale Hypothesenklassen gültig ist:

Theorem 6.6 *Für alle nicht-trivialen binären Hypothesenklassen $\mathcal{H}$ und alle $0 < \varepsilon \leq 1/4$, $0 < \delta < 1$ gilt (sogar unter der Realisierbarkeitsannahme):*

$$m_{\mathcal{H}}(\varepsilon, \delta) > \frac{1}{4\varepsilon} \ln\left(\frac{1}{4\delta}\right)$$

Beweis Wähle $x_1, x_2 \in \mathcal{X}$ und $h_1, h_2 \in \mathcal{H}$ so aus, dass die Bedingung (4) erfüllt ist. Wähle $\mathcal{D}$ bzw. P so, dass die gesamte Wahrscheinlichkeitsmasse auf x_1, x_2 bzw. h_1, h_2 konzentriert ist, wobei $\mathcal{D}(x_1) = 1 - 2\varepsilon$, $\mathcal{D}(x_2) = 2\varepsilon$ und $P(h_1) = P(h_2) = 1/2$. Es sei B_1 das Ereignis, dass $S|_{\mathcal{X}} \sim \mathcal{D}^m$ aus m Duplikaten von x_1 besteht. Wegen $h_1(x_1) = h_2(x_1)$ unterscheidet die resultierende Trainingsmenge S dann nicht zwischen den möglichen Labelfunktionen h_1 und h_2. Wegen $h \sim P$ und der Wahl von P ist $h(x_2)$ aus Sicht des Lerners ein Zufallsbit. Es sei B_2 das Ereignis $h_S(x_2) \neq h(x_2)$. Wenn B_2 eintritt, dann gilt $L_{\mathcal{D},h}(h_S) = 2\varepsilon$. Wir wollen im folgenden zeigen, dass

$$m \leq \frac{1}{4\varepsilon} \ln\left(\frac{1}{4\delta}\right)$$

eine zu kleine Wahl für die Größe der Trainingsmenge ist, Mit dieser Wahl von m ist die Wahrscheinlichkeit von B_1 gleich $(1 - 2\varepsilon)^m \geq e^{-4\varepsilon m} \geq 4\delta$. Die bedingte Wahrscheinlichkeit von B_2 zu gegebenem B_1 beträgt $1/2$. Also treten mit einer Wahrscheinlichkeit von mindestens 2δ beide Ereignisse (insbesondere B_2) ein. Es gilt dann

$$\Pr_{S|_{\mathcal{X}} \sim \mathcal{D}^m, h \sim P}[L_{\mathcal{D},h}(h_S)] \geq 2\varepsilon] \geq 2\delta$$

und das Erfolgskriterium (2) wird verfehlt. **qed.**

Das folgende Resultat zeigt, dass $m_{\mathcal{H}}(\varepsilon, \delta)$ (sogar unter der Realisierbarkeitsannahme) mindestens linear mit $d = \text{VCdim}(\mathcal{H})$ wachsen muss.

Theorem 6.7 *Für alle binären Hypothesenklassen und alle $0 < \varepsilon \leq 1/8$, $0 < \delta < 1/4$ gilt (sogar unter der Realisierbarkeitsannahme)*

$$m_{\mathcal{H}}(\varepsilon, \delta) > \frac{d-1}{32\varepsilon} \quad .$$

Beweis Es sei $t = d - 1$ und $\mathcal{X}_0 = \{x_0, x_1, \ldots, x_t\} \subseteq \mathcal{X}$ sei eine von $\mathcal{H}$ aufspaltbare Menge.[2] Weiter sei $\mathcal{X}_0' = \mathcal{X}_0 \backslash \{x_0\}$. Die Verteilung $\mathcal{D}$ konzentriere ihre ganze Wahrscheinlichkeitsmasse auf $\mathcal{X}_0$ und es gelte

$$\mathcal{D}(x_0) = 1 - 8\varepsilon \quad \text{und} \quad \mathcal{D}(x_1) = \ldots = \mathcal{D}(x_t) = \frac{8\varepsilon}{t} \ .$$

Wähle eine Menge $\mathcal{H}_0$ aus 2^t Hypothesen aus, welche die 2^t Bitmuster $b = (b_0, b_1, \ldots, b_t)$ mit $b_0 = 0$ realisieren. Die Verteilung P konzentriere ihre ganze Wahrscheinlichkeitsmasse auf $\mathcal{H}_0$ und sei dort uniform. Im Folgenden identifizieren wir die Hypothesen aus $\mathcal{H}_0$ mit den von ihnen realisierten Bitmustern b. Beachte, dass eine zufällige Wahl von $h \sim P$ äquivalent dazu ist, dass das von h realisierte Bitmuster b (abgesehen von $b_0 = 0$) aus t unabhängigen perfekten Zufallsbits besteht. Betrachte nun eine Menge $M = S|_{\mathcal{X}} \sim \mathcal{D}^m$ von unmarkierten Trainingspunkten. Es sei X die Zufallsvariable, welche zählt wieviele Treffer M in $\mathcal{X}_0'$ landet. Es sei B_1 das Ereignis $X \leq t/2$. Aus Sicht des Lerners sind die Labels der nicht in M enthaltenen Instanzen aus $\mathcal{X}_0'$ unabhängige perfekte Zufallsbits. Es sei B_2 das Ereignis, dass h_S auf $\mathcal{X}_0'$ mindestens $t/4$ Fehler macht. Wenn das Ereignis B_2 eintritt, dann gilt offensichtlich $L_{\mathcal{D},h}(h_S) \geq 2\varepsilon$ (da $t/4$ Punkte aus $\mathcal{X}_0'$ eine Wahrscheinlichkeitsmasse von insgesamt 2ε besitzen). Wir bestimmen abschließend eine untere Schranke für die Wahrscheinlichkeit von $B_1 \wedge B_2$. Unter der Voraussetzung

$$m \leq \frac{t}{32\varepsilon}$$

gilt $\mathbb{E}[X] \leq t/4$. Aus der Markov Ungleichung folgt, dass die Wahrscheinlichkeit von $X > t/2$ (das Ereignis $\neg B_1$) höchstens $1/2$ beträgt. Die Wahrscheinlichkeit von B_1 ist daher mindestens $1/2$. Die bedingte Wahrscheinlichkeit von B_2 gegeben B_1 beträgt mindestens $1/2$.[3] Es folgt, dass beide Ereignisse (insbesondere B_2) mit einer Wahrscheinlichkeit von mindestens $1/4$ auftreten. Es gilt daher

$$\Pr_{S|_{\mathcal{X}} \sim \mathcal{D}^m, h \sim P}[L_{\mathcal{D},h}(h_S) \geq 2\varepsilon] \geq \frac{1}{4}$$

und das Erfolgskriterium (2) wird verfehlt. **qed.**

Aus Theorem 6.7 können wir unmittelbar die Nichtlernbarkeit von Klassen unendlicher VC-Dimension ableiten:

Folgerung 6.8 *Eine binäre Hypothesenklasse unendlicher Dimension ist nicht PAC-lernbar, und zwar nicht einmal, wenn wir die Realisierbarkeitsannahme machen und zudem ε konstant auf $1/8$ und δ konstant auf $1/5$ setzen.*

[2] Der folgende Beweis unterstellt implizit, dass t eine durch 4 teilbare Zahl ist. Er lässt sich aber so modifizieren, dass er auch für andere Werte von t funktioniert.

[3] Genau hier verwenden wir die Teilbarkeit von t durch 4.

6.4 Sauer's Lemma und die Kapazitätsfunktion

Eine zentrale Definition in der Lerntheorie ist die folgende:

Definition 6.9 (Kapazitätsfunktion) *Die* Kapazitätsfunktion $\tau_{\mathcal{H}}(m)$ *liefert die maximale Anzahl von Bitmustern, die von $\mathcal{H}$ auf einer m-elementigen Teilmenge von $\mathcal{X}$ realisiert werden können. Formal:*

$$\tau_{\mathcal{H}}(m) = \sup\{|\mathcal{H}_M| : M \subseteq \mathcal{X} \wedge |M| = m\} \ .$$

Trivialerweise gilt $\tau_{\mathcal{H}}(m) \leq 2^m$. Dieser Maximalwert wird genau dann erreicht, wenn m die VC-Dimension von $\mathcal{H}$ nicht überschreitet. Eine exponentiell mit m wachsende Zahl von durch $\mathcal{H}$ realisierbaren Bitmustern der Länge m, würde PAC-Lernbarkeit von $\mathcal{H}$ ausschließen. Glücklicherweise — und das wird das Hauptresultat dieses Abschnittes — wächst $\tau_{\mathcal{H}}(m)$ nur polynomiell in Hypothesenklassen $\mathcal{H}$ mit endlicher VC-Dimension d. Genauer: $\tau_{\mathcal{H}}(m) = O(m^d)$, wie wir im Folgenden noch zeigen werden.

Lemma 6.10 *Es sei $\mathcal{H}$ eine binäre Hypothesenklasse über einem Grundbereich M der Größe $m \geq 1$. Dann ist $|\mathcal{H}|$ nach oben beschränkt durch die Anzahl der von $\mathcal{H}$ aufspaltbaren Teilmengen von M.*

Beweis Der Beweis erfolgt mit vollständiger Induktion. Für $m = 1$ ist die Aussage offensichtlich. Sei nun $m \geq 2$, x ein beliebiges aus M ausgewähltes Element und $M' = M \setminus \{x\}$. Wir definieren folgende Hypothesenklassen $\mathcal{H}_b$, $b = 0, 1$, über dem Grundbereich M':

$$\mathcal{H}_b = \{h : M' \to \{0,1\} |\ h \in \mathcal{H} \wedge h(x) = b\} \ .$$

Es gilt $|\mathcal{H}| = |\mathcal{H}_0| + |\mathcal{H}_1|$, da $\mathcal{H}$ disjunkt in $\mathcal{H}_0$ und $\mathcal{H}_1$ zerlegt werden kann. Es bezeichne a die Anzahl der von $\mathcal{H}$ aufspaltbaren Teilmengen von M und a_b sei die entsprechende Zahl für $\mathcal{H}_b$. Mit der Induktionsvoraussetzung folgt $|cH_b| \leq a_b$ für $b = 0, 1$. Es gilt $a \geq a_0 + a_b$, und zwar aus folgenden Gründen:

- Jede Menge, die von $\mathcal{H}_0$ oder $\mathcal{H}_1$ aufspaltbar ist, ist auch von $\mathcal{H}$ aufspaltbar.
- Der Term $a_0 + a_1$ zählt zwar jede Mengen $A \subseteq M'$ doppelt, die sowohl von $\mathcal{H}_0$ als auch von $\mathcal{H}_1$ aufspaltbar ist, aber dann ist nicht nur A von $\mathcal{H}$ aufspaltbar sondern auch $A \cup \{x\}$.

Insgesamt hat sich $|\mathcal{H}| = |\mathcal{H}_0| + |\mathcal{H}_1| \leq a_0 + a_1 \leq a$ ergeben, was den induktiven Beweis abschließt. **qed.**

Die folgende Funktion liefert (wie wir gleich zeigen werden) eine obere Schranke für die Werte der Kapazitätsfunktion zu einer Hypothesenklasse der VC-Dimension d:

$$\Phi_d(m) = \sum_{i=0}^{d} \binom{m}{d} \leq \left(\frac{em}{d}\right)^d \ .$$

Lemma 6.11 (Sauer's Lemma) *Es sei $\mathcal{H}$ eine binäre Hypothesenklasse der VC-Dimension d. Dann gilt $\tau_{\mathcal{H}}(m) \leq \Phi_d(m)$.*

Beweis Es genügt $|\mathcal{H}_M| \leq \Phi_d(m)$ für eine beliebige Menge $M \subseteq \mathcal{X}$ mit $|M| = m$ nachzuweisen. Gemäß Lemma 6.10 ist $|\mathcal{H}_M|$ nach oben beschränkt durch die Anzahl der von $\mathcal{H}$ aufspaltbaren Teilmengen von M. Wegen $d = \mathrm{VCdim}(\mathcal{H})$ hat keine dieser aufspaltbaren Teilmengen mehr als d Elemente. Da $\Phi_d(m)$ mit der Anzahl aller Teilmengen von M mit höchstens d Elementen übereinstimmt, folgt $\tau_{\mathcal{H}}(m) \leq \Phi_d(m)$. **qed.**

Die Schranke in Sauer's Lemma ist gelegentlich scharf, zum Beispiel für die Klasse aller Teilmengen von $\mathbb{N}$ mit höchstens d Elementen. Dies ist eine Klasse der VC-Dimension d, deren Kapazitätsfunktion mit $\Phi_d(m)$ übereinstimmt.

6.5 Obere Schranke zur Informationskomplexität

In diesem Abschnitt bezeichnen wir mit Q die Menge der für $\mathcal{H}$ nicht ε-repräsentativen Trainingsmengen der Größe m. Formal:

$$Q = \{z \in Z^m | \ \exists h \in \mathcal{H} : |L_z(h) - L_{\mathcal{D}}(h)| > \varepsilon\} \ .$$

Wir wollen in diesem Abschnitt nachweisen, dass

$$\mathcal{D}^m(Q) \leq 4\tau_{\mathcal{H}}(2m) \exp\left(-\frac{\varepsilon^2 m}{8}\right) \ , \tag{6}$$

vorausgesetzt dass $m \geq 2\ln(4)/\varepsilon^2$. Der Nachweis ist nichttrivial und bedarf einiger begrifflicher Vorbereitungsarbeiten.

Eine Trainingsmenge $z \in Z^{2m}$ der Größe $2m$ hat die Form $z = (z_1, \ldots, z_{2m})$ mit $z_i = (x_i, y_i) \in Z = \mathcal{X} \times \{0, 1\}$. Wir zerlegen z in Gedanken in zwei Trainingsmengen z', z'' der Größe m, so dass $z = (z', z'')$ mit $z' = (z_1, \ldots, z_m)$ und $z'' = (z_{m+1}, \ldots, z_{2m})$. Es wird sich als zweckmäßig herausstellen, die Menge Q in Beziehung zu setzen zu folgender Menge:

$$\begin{aligned} R &= \{z \in Z^m \times Z^m | \ \exists h \in \mathcal{H} : |L_{z'}(h) - L_{z''}(h)| \geq \varepsilon/2\} \\ &= \left\{ z \in Z^{2m} | \ \exists h \in \mathcal{H} : \frac{1}{m} \left| \sum_{i=1}^{m} \mathbb{1}_{[h(x_i) \neq y_i]} - \mathbb{1}_{[h(x_{m+i}) \neq y_{m+i}]} \right| \geq \varepsilon/2 \right\} \ . \end{aligned}$$

Die Definitionsbedingung der Menge R in Worten lautet: es gibt eine Hypothese in $\mathcal{H}$, bei welcher die empirischen Fehlerraten bezüglich z' und z'' um mindestens $\varepsilon/2$ von einander abweichen. Für $b \in \{\pm 1\}^m$ definieren wir folgende zu R verwandten Ereignisse:

$$\begin{aligned} R_b &= \left\{ z \in Z^{2m} | \ \exists h \in \mathcal{H} : \frac{1}{m} \left| \sum_{i=1}^{m} b_i \cdot \left(\mathbb{1}_{[h(x_i) \neq y_i]} - \mathbb{1}_{[h(x_{m+i}) \neq y_{m+i}]} \right) \right| \geq \varepsilon/2 \right\} \\ R_{b,h} &= \left\{ z \in Z^{2m} | \ \frac{1}{m} \left| \sum_{i=1}^{m} b_i \cdot \left(\mathbb{1}_{[h(x_i) \neq y_i]} - \mathbb{1}_{[h(x_{m+i}) \neq y_{m+i}]} \right) \right| \geq \varepsilon/2 \right\} \ . \end{aligned}$$

Der Nachweis der oberen Schranke (6) benutzt folgende (noch zu verifizierende) Zwischenschritte:

Behauptung 1: $\mathcal{D}^m[Q] \leq 2\mathcal{D}^{2m}(R)$, sofern $m \geq 2\ln(4)/\varepsilon^2$.

Behauptung 2: Für alle $b \in \{\pm 1\}^m$ gilt: $D^{2m}(R) = D^{2m}(R_b)$.

Behauptung 3: Es bezeichne U die uniforme Verteilung auf $\{\pm 1\}^m$. Dann gilt:

$$\mathcal{D}^{2m}(R) \leq \sup_{z \in Z^{2m}} \Pr_{b \sim U}[z \in R_b] \ .$$

Behauptung 4: Für jede feste Wahl von $z \in Z^{2m}$ gilt:

$$\Pr_{b \sim U}[z \in R_b] \leq \tau_{\mathcal{H}}(2m) \sup_{h \in \mathcal{H}} \Pr_{b \sim U}[z \in R_{b,h}] \ .$$

Behauptung 5: Für jede feste Wahl von $z \in Z^{2m}$ und $h \in \mathcal{H}$ gilt:

$$\Pr_{b \sim U}[z \in R_{b,h}] \leq 2\exp\left(-\frac{\varepsilon^2 m}{8}\right) \ .$$

Aus diesen fünf Behauptungen ist (6) direkt ersichtlich. Es folgen die Nachweise, dass die aufgestellten Behauptungen korrekt sind.

Beweis von Behauptung 1: Für jedes $z \in Q$ wählen wir eine Funktion $h_z \in \mathcal{H}$ aus mit $|L_z(h_z) - L_{\mathcal{D}}(h_z)| > \varepsilon$. Wir starten mit folgender Rechnung:

$$\begin{aligned}
\mathcal{D}^{2m}(R) &= \Pr_{z \sim \mathcal{D}^{2m}}[z \in R] \\
&\geq \Pr_{z \sim \mathcal{D}^{2m}}[z \in R \wedge z' \in Q] \\
&= \Pr_{z \sim \mathcal{D}^{2m}}[(z', z'') \in R | z' \in Q] \cdot \Pr_{z' \sim \mathcal{D}^m}[z' \in Q] \\
&= \Pr_{z \sim \mathcal{D}^{2m}}[(z', z'') \in R | z' \in Q] \cdot \mathcal{D}^m(Q) \ .
\end{aligned}$$

Zum Nachweis von Behauptung 1 ist noch $\Pr_{z \sim \mathcal{D}^{2m}}[(z', z'') \in R | z' \in Q] \geq 1/2$ zu zeigen. Wir stellen folgende Überlegung an. Gegeben dass $z' \in Q$, also dass $L_{z'}(h)$ um mehr als ε von $L_{\mathcal{D}}(h)$ abweicht, dann kann das Ereignis $(z', z'') \notin R$ nur dann eintreten, wenn für alle Hypothesen $h \in \mathcal{H}$ die Bedingung $|L_{z'}(h) - L_{z''}(h)| < \varepsilon/2$ erfüllt ist. Insbesondere muss dies für $h_{z'}$ gelten. Wegen $|L_{z'}(h_{z'}) - L_{\mathcal{D}}(h_{z'})| > \varepsilon$ folgt dann zwangsläufig, dass $|L_{z''}(h_{z'}) - L_{\mathcal{D}}(h_{z'})| > \varepsilon/2$. Gemäß der Hoeffding-Schranke beträgt die Wahrscheinlichkeit für letzteres Ereignis höchstens $2\exp(-2m(\varepsilon/2)^2) = 2\exp(-m\varepsilon^2/2) \leq 1/2$, wobei die letzte Ungleichung wegen $m \geq 2\ln(4)/\varepsilon^2$ gültig ist. Aus dieser Diskussion folgt, $\Pr_{z \sim \mathcal{D}^{2m}}[(z', z'') \notin R | z' \in Q] \leq 1/2$ und somit $\Pr_{z \sim \mathcal{D}^{2m}}[(z', z'') \in R | z' \in Q] \geq 1/2$. Damit ist der Beweis von Behauptung 1 komplett.

Beweis von Behauptung 2: Für jede Permutation σ von $1, \ldots, 2m$ sind die Zufallsvariablen $z = (z_1, \ldots, z_{2m}) \sim \mathcal{D}^{2m}$ und $z^\sigma = (z_{\sigma(1)}, \ldots, z_{\sigma(2m)})$ mit $z \sim \mathcal{D}^{2m}$ gleichverteilt. Daher gilt $\Pr_{z\sim\mathcal{D}^{2m}}[z \in R] = \Pr_{z\sim\mathcal{D}^{2m}}[z^\sigma \in R]$. Zu $b \in \{\pm 1\}^m$ betrachte nun die Permutation σ_b, die (für $i = 1, \ldots, m$) i mit $m + i$ vertauscht, falls $b_i = -1$, und $i, m + i$ als Fixpunkte hat, falls $b_i = 1$. Offensichtlich gilt $z^{\sigma_b} \in R \Leftrightarrow z \in R_b$. Hieraus ergibt sich die Behauptung 2.

Beweis von Behauptung 3: Die Behauptung ergibt sich aus folgender Rechnung:

$$\begin{aligned} \mathcal{D}^{2m}(R) &= 2^{-m} \sum_{b\in\{\pm1\}^m} D^{2m}(R_b) = \mathbb{E}_{b\sim U, z\sim\mathcal{D}^{2m}}[\mathbb{1}_{[z\in R_b]}] \\ &\leq \sup_{z\in Z^{2m}} \mathbb{E}_{b\sim U}[\mathbb{1}_{[z\in R_b]}] = \sup_{z\in Z^{2m}} \Pr_{b\sim U}[z \in R_b] \ . \end{aligned}$$

Beweis von Behauptung 4: Für eine feste Wahl von $z \in Z^{2m}$ spielen zur Berechnung von $\Pr_{b\sim U}[z \in R_b]$ bei jeder Hypothese $h \in \mathcal{H}$ nur ihre Werte auf $M = z|_\mathcal{X}$ eine Rolle. Diese Wahrscheinlichkeit würde sich nicht ändern, wenn wir auf der rechten Seite der Definitionsgleichung für R_b "$\exists h \in \mathcal{H}_M$" statt "$\exists h \in \mathcal{H}$" schreiben. Aus der Definition der Kapazitätsfunktion $\tau_\mathcal{H}$ folgt $|\mathcal{H}_M| \leq \tau_\mathcal{H}(2m)$. Behauptung 4 ergibt sich daher aus der „Union Bound".

Beweis von Behauptung 5: Für feste Wahl von z, h und $b \sim U$ betrachte die Zufallsvariablen

$$X_i = b_i \cdot (\mathbb{1}_{[h(x_i)\neq y_i]} - \mathbb{1}_{[h(x_{m+i})\neq y_{m+i}]}) \in \{-1, 0, 1\} \subset [-1, 1]$$

für $i = 1, \ldots, m$. Dann sind $X_1, \ldots, X_m$ iid-Variablen mit Erwartungswert 0. Die Aussage $z \in R_{b,h}$ ist äquivalent dazu, dass $(X_1 + \ldots + X_m)/m$ um mindestens $\varepsilon/2$ vom Erwartungswert 0 abweicht. Mit der Hoeffding-Ungleichung erhalten wir $\Pr_{b\sim U}[z \in R_{b,h}] \leq 2\exp(-\varepsilon^2 m/8)$, wie man leicht nachrechnen könnte.

Wir wollen rückblickend noch einmal auf wichtige Techniken aufmerksam machen, die beim Nachweis von (6) zur Anwendung kamen. Behauptungen 2 und 3 bringen die Vorzeichenmuster $b \in \{\pm 1\}^m$ ins Spiel und nutzen aus, dass $\mathcal{D}^{2m}(R_b)$ für alle Wahlen von b stets den selben Wert hat (eine Art Symmetrie weswegen man auch von einer „Symmetrisierungstechnik" spricht). Ab Behauptung 3 haben wir es mit einem endlichen Wahrscheinlichkeitsraum mit den Elementarereignissen $b \in \{\pm 1\}^m$ zu tun. Dieser ist wesentlich einfacher zu analysieren als der ursprüngliche Raum mit der (dem Lerner unbekannten) Verteilung $\mathcal{D}^{2m}$. Beim Übergang von Behauptung 3 zu Behauptung 4 werden wir den Existenzquantor los, der in den Definitionsbedingungen der Ereignisse R und R_b steckt. Um ihn loszuwerden, haben wir Sauer's Lemma und die Union Bound eingesetzt. Danach ist das Problem soweit „runter gekocht", dass (im Beweis von Behauptung 5) schließlich die Hoeffding Ungleichung zum Einsatz gebracht werden kann.

Wenn wir (6) mit dem Lemma von Sauer kombinieren, erhalten wir:

Theorem 6.12 *Es sei $\mathcal{H}$ eine binäre Hypothesenklasse der VC-Dimension d. Dann gilt für alle $m \geq 2\ln(4)/\varepsilon^2$:*

$$\Pr_{z\sim\mathcal{D}^m}[\exists h \in \mathcal{H} : |L_z(H) - L_{\mathcal{D}}(\mathcal{H})| > \varepsilon] \leq 4\Phi_d(2m)\exp\left(\frac{-\varepsilon^2 m}{8}\right) \leq 4\left(\frac{2em}{d}\right)^d \exp\left(\frac{-\varepsilon^2 m}{8}\right) . \tag{7}$$

Weiterhin existiert eine universelle Konstante $c > 0$, so dass

$$m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \leq c \cdot \frac{1}{\varepsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right) .$$

Insbesondere erfüllt jede binäre Hypothesenklasse einer endlichen VC-Dimension die uniforme Konvergenzbedingung.

Beweis Die erste Aussage ist klar. Die zweite ergibt sich aus der ersten, indem wir die rechte Seite von (7) kleiner oder gleich δ setzen und dann (geschickt) nach m auflösen. **qed.**

6.6 Das Fundamentaltheorem der PAC-Lerntheorie

Die qualitative Version des Fundamentaltheorems liest sich wie folgt:

Theorem 6.13 *Für eine binäre Hypothesenklasse (und die Null-Eins-Verlustfunktion) sind folgende Aussagen äquivalent:*

1. *$\mathcal{H}$ erfüllt die uniforme Konvergenzbedingung.*
2. *$\mathcal{H}$ ist mit ERM agnostisch PAC-lernbar.*
3. *$\mathcal{H}$ ist agnostisch PAC-lernbar.*
4. *$\mathcal{H}$ ist PAC-lernbar unter der Realisierbarkeitsannahme.*
5. *$\mathcal{H}$ ist mit ERM PAC-lernbar unter der Realisierbarkeitsannahme.*
6. *Die VC-Dimension von $\mathcal{H}$ ist endlich.*

Beweis Aus dem Erfüllen der uniformen Konvergenzbedingung ergeben sich alle Aussagen zur PAC-Lernbarkeit von $\mathcal{H}$, also die Aussagen 2,3,4,5. Aus jedem der PAC-Lernbarkeitsresultate ergibt sich die Endlichkeit der VC-Dimension, da gemäß Folgerung 6.8 Klassen mit unendlicher VC-Dimension nicht einmal unter der Realisierbarkeitsannahme PAC-lernbar sind. Zu einem vollständigen Ringschluss fehlt also nur die Herleitung der ersten Aussage (Erfüllen der uniformen Konvergenzbedingung) aus der sechsten (Endlichkeit der VC-Dimension). Diesen fehlenden Puzzlestein liefert aber gerade Theorem 6.12. **qed.**

Es folgt die quantitative Version des Fundamentaltheorems:

Theorem 6.14 *Es gibt universelle Konstanten $c_1, \ldots, c_7 > 0$, so dass für jede binäre Hypothesenklasse und $d = VCdim(\mathcal{H})$ folgendes gilt:*

$$\begin{aligned} c_1 \cdot \frac{1}{\varepsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right) &\leq m_{\mathcal{H}}^{UC}(\varepsilon,\delta) \leq c_2 \cdot \frac{1}{\varepsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right) \\ c_3 \cdot \frac{1}{\varepsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right) &\leq m_{\mathcal{H}}(\varepsilon,\delta) \leq c_4 \cdot \frac{1}{\varepsilon^2}\left(d + \ln\left(\frac{1}{\delta}\right)\right) \end{aligned}$$

Unter der Realisierbarkeitsannahme gilt weiterhin

$$c_5 \cdot \frac{1}{\varepsilon}\left(d + \ln\left(\frac{1}{\delta}\right)\right) \leq m_{\mathcal{H}}(\varepsilon,\delta) \leq c_6 \cdot \frac{1}{\varepsilon}\left(d\log\left(\frac{1}{\varepsilon}\right) + \ln\left(\frac{1}{\delta}\right)\right)$$

bzw. sogar

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq c_7 \cdot \frac{1}{\varepsilon}\left(d + \ln\left(\frac{1}{\delta}\right)\right) \quad ,$$

wenn der Lernalgorithmus eine Hypothese außerhalb von $\mathcal{H}$ ausgeben darf. Alle oberen Schranken mit Ausnahme der letzten werden durch die ERM-Lernstrategie (Ausgabe einer Hypothese aus $\mathcal{H}$ mit minimalem empirischen Verlustwert) erreicht.

Theorem 6.14 haben wir nur teilweise bewiesen:

- Die angegebene obere Schranke für $m_{\mathcal{H}}^{UC}(\varepsilon,\delta)$ wurde im Rahmen von Theorem 6.12 bewiesen.
- Die angegebene obere Schranke für $m_{\mathcal{H}}(\varepsilon,\delta)$ folgt dann aus der allgemeinen Ungleichung $m_{\mathcal{H}}(\varepsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2,\delta)$.
- Die unter der Realisierbarkeitsannahme angegebene untere Schranke für $m_{\mathcal{H}}(\varepsilon,\delta)$ haben wir in Abschnitt 6.3 hergeleitet.

Wir verzichten auf den vollständigen Beweis.