AGML, Sommersemester 2020 Vorlesung 5 (4 Seiten)

Mehr zum Nicht-Realisierbaren Fall und Grenzen der Lernbarkeit

Thomas Kesselheim **Vorschau** Letzte Aktualisierung: 8. Mai 2020

In der vergangenen Vorlesung haben wir die Definition von PAC-Lernen mit agnostischem Sinn kennengelernt. Hier gibt es eine Verteilung $\mathcal{D}$ über Datenpunkt-/Label-Paaren, also über der Menge $X \times \{-1, +1\}$. Der tatsächliche Fehler einer Hypothese h is definiert als

$$\mathrm{err}_{\mathcal{D}}(h) := \mathbf{Pr}_{(x,y)\sim\mathcal{D}}\left[h(x) \neq y\right] \ .$$

Es gibt im Allgemeinen keine Grundwahrheit f, die eine mögliche Hypothese ist. In diesem Fall gilt auch $\min_{h'\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h') > 0$. Es ist somit nicht möglich, dass der tatsächliche Fehler eines Algorithmus verschwindet, egal wie viele Samples wir ihm bereitstellen. Stattdessen ist das Ziel, möglichst nah an $\min_{h'\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h')$ heranzukommen.

1 Minimieren des Trainingsfehlers im Agnostischen Fall

Gegeben eine Trainingsmenge $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ können wir den Trainingsfehler eine Hypothese definieren als

$$\mathrm{err}_S(h) := \frac{1}{m}|\{h(x_i) \neq y_i\}| \ .$$

Wir können uns nun Algorithmen anschauen, die diesen Trainingsfehler minimieren. Während dies im realisierbaren Fall bedeutet, dass kein Fehler auf S gemacht werden darf, ist dies nun nicht immer möglich. Es ist nur das Ziel, möglichst wenige Fehler zu machen.

Für den agnostischen Fall kann man eine sehr ähnliche Aussage wie im realisierbaren Fall herleiten, die die Wachstumsfunktion nutzt.

Satz 5.1. *Seien eine $\mathcal{H}$ beliebige Hypothesenklasse über X und $\mathcal{D}$ eine Verteilung über $X \times \{-1, +1\}$. Seien $\epsilon > 0$, $\delta > 0$ beliebig und*

$$m \geq \frac{32}{\epsilon^2} \ln\left(\frac{4\Pi_{\mathcal{H}}(2m)}{\delta}\right) \ .$$

Betrachte ein Sample S von m Datenpunkten mit Labels gezogen unabhängig und identisch verteilt aus $\mathcal{D}$. Es gilt mit Wahrscheinlichkeit mindestens $1 - \delta$, dass jede Hypothese h, die $\mathrm{err}_S(h)$ minimiert, auch $\mathrm{err}_{\mathcal{D}}(h) \leq \min_{h'\in\mathcal{H}} \mathrm{err}_{\mathcal{D}}(h') + \epsilon$ erfüllt.

Insbesondere folgt aus dieser Schranke auch, dass eine Hypothesenklassen im agnostischen Sinn PAC-lernbar ist, wenn ihr VC-Dimension endlich ist. Der Lernalgorithmus ist in diesem Fall ein beliebiger Algorithmus, der den Trainingsfehler minimiert.

Viele Schritte im Beweis dieses Satzes sind analog zu seinem Pendant im realisierbaren Fall. Um die Unterschiede und zusätzlichen Techniken zu verdeutlichen, betrachten wir nun den Fall einer *endlichen* Hypothesenklasse $\mathcal{H}$. Wir zeigen, dass für

$$m \geq \frac{2}{\epsilon^2} \ln\left(\frac{2|\mathcal{H}|}{\delta}\right) \tag{1}$$

die Aussage von Satz 5.1 erfüllt ist. Hierzu beweisen wir folgende Behauptung.

Behauptung 5.2.

$$\mathbf{Pr}\left[\exists h \in \mathcal{H} : |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \frac{\epsilon}{2}\right] < \delta \ .$$

Diese Aussage hilft uns wie folgt. Angenommen, wir haben eine Menge S, sodass

$$|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| < \frac{\epsilon}{2} \qquad \text{für alle } h \in \mathcal{H}. \tag{2}$$

Das heißt, der tatsächliche Fehler und der Trainingsfehler sind nah bei einander für jede mögliche Hypothese. Ist nun h eine Hypothese, die den Trainingsfehler $\mathrm{err}_S(h)$ minimiert; h' eine Hypothese, die den tatsächlichen Fehler $\mathrm{err}_{\mathcal{D}}(h')$ minimiert, dann gilt

$$\mathrm{err}_{\mathcal{D}}(h) < \mathrm{err}_S(h) + \frac{\epsilon}{2} \leq \mathrm{err}_S(h') + \frac{\epsilon}{2} < \mathrm{err}_{\mathcal{D}}(h') + \epsilon \ .$$

Für den Beweis von Behauptung 5.2 zeigen nun wieder zunächst eine Aussage über eine einzelne Hypothese.

Lemma 5.3. *Betrachte eine feste Hypothese $h \in \mathcal{H}$. Sei S eine Menge von m Datenpunkt-/Label-Paaren aus $\mathcal{D}$. Dann gilt für alle $\gamma > 0$*

$$\mathbf{Pr}\left[|\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \gamma\right] \leq 2\exp\left(-2m\gamma^2\right) \ .$$

Beweis. Diese Aussage folgt einigermaßen direkt aus der Hoeffding-Ungleichung. Diese lautet wie folgt.

Lemma 5.4 (Hoeffding-Ungleichung)**.** *Seien $Z_1, \ldots, Z_N$ unabhängige Zufallsvariablen, sodass $a_i \leq Z_i \leq b_i$ mit Wahrscheinlichkeit 1. Sei $\bar{Z} = \frac{1}{N}\sum_{i=1}^{N} Z_i$ ihr Durchschnitt. Dann gilt für alle $\gamma \geq 0$*

$$\mathbf{Pr}\left[|\bar{Z} - \mathbf{E}\left[\bar{Z}\right]| \geq \gamma\right] \leq 2\exp\left(-\frac{2N^2\gamma^2}{\sum_{i=1}^{N}(b_i - a_i)^2}\right) \ .$$

Die Ungleichung quantifiziert (und verallgemeinert) das Gesetz der großen Zahlen: Der Durchschnitt vieler Züge aus derselben Verteilung konvergiert gegen den Erwartungswert.

Für unsere Aussage sei $Z_i = 1$, falls $h(x_i) \neq y_i$ und 0 sonst. Dann gilt $\bar{Z} = \mathrm{err}_S(h)$. Außerdem sind $Z_1, \ldots, Z_m$ unabhängig und es gilt $0 \leq Z_i \leq 1$. Also können wir die Hoeffing-Ungleichung mit $a_i = 0$, $b_i = 1$ and $N = m$ anwenden.

Schließlich stellen wir fest, dass $\mathbf{E}\left[Z_i\right] = \mathrm{err}_{\mathcal{D}}(h)$ für alle i und damit auch $\mathbf{E}\left[\bar{Z}\right] = \frac{1}{m}\sum_{i=1}^{m}\mathbf{E}\left[Z_i\right] = \mathrm{err}_{\mathcal{D}}(h)$. Die Aussage des Lemmas ist also genau die Schranke, die aus der Hoeffding-Ungleichung folgt. □

Jetzt ist der Beweis von Behauptung 5.2 auch unkompliziert.

Beweis von Behauptung 5.2. Wir nutzen wieder die Union Bound and wählen $\gamma = \frac{\epsilon}{2}$ in Lemma 5.3. Damit bekommen wir

$$\mathbf{Pr}\left[\exists h \in \mathcal{H} : |\mathrm{err}_{\mathcal{D}}(h) - \mathrm{err}_S(h)| \geq \frac{\epsilon}{2}\right] \leq |\mathcal{H}| \cdot 2\exp\left(-m\frac{\epsilon}{2}\right) \leq \delta \ .$$ □

2 Unendliche VC-Dimension

Wir haben bereits gesehen, dass jede Hypothesenklassen $\mathcal{H}$ endlicher VC-Dimension PAC-lernbar ist. Aber was ist im Fall von unendlicher VC-Dimension? Beispielsweise die Klasse aller Hypothesen $\mathbb{N} \to \{-1, +1\}$. Oder allgemeiner alle Funktionen $X \to \{-1, +1\}$. Wie wir zeigen werden, sind diese nicht PAC-lernbar.

Satz 5.5. *Jede Hypothesenklasse von unendlicher VC-Dimension ist nicht PAC-lernbar im realisierbaren Sinn.*

Um diesen Satz zu beweisen, müssen wir zeigen, dass Lernalgorithmus $\mathcal{A}$ und Funktion $m_{\mathcal{H}}$ aus der Definition von PAC-Lernbarkeit nicht existieren. Wir werden die folgende Aussage zeigen.

Behauptung 5.6. *Sei $\mathcal{H}$ eine Hypothesenklasse von VC-Dimension mindestens d. Dann gibt es für jeden Lernalgorithmus $\mathcal{A}$ eine Verteilung $\mathcal{D}$ und eine Grundwahrheit f, sodass auf einer Trainingsmenge S der Größe höchstens $\frac{d}{2}$ gilt:* $\operatorname{err}_{\mathcal{D}}(h_S) \geq \frac{1}{8}$ *mit Wahrscheinlichkeit mindestens* $\frac{1}{7}$.

Beweis. Laut Definition spaltet $\mathcal{H}$ eine Menge der Größe d auf. Sei also $T \subseteq X$, $|T| = d$, eine solche Menge. Es gilt nun $|\mathcal{H}|_T| = 2^d$. Definiere $k = 2^d$ und schreibe $\mathcal{H}|_T = \{\ell_1, \ldots, \ell_k\}$, wobei jeweils $\ell_i \colon T \to \{-1, +1\}$ und alle ℓ_i unterschiedlich sind.

Für jedes ℓ_i finden wir ein $f_i \in \mathcal{H}$, sodass $f_i(x) = \ell_i(x)$ für alle $x \in X$. Jede dieser Funktionen f_i könnte die Grundwahrheit sein. Die entscheidende Beobachtung ist, dass wenn uns lediglich ein Sample der Größe $\frac{d}{2}$ gegeben wird, wir für höchstens $\frac{d}{2}$ Punkte in T das korrekte Label wissen. Für die übrigen Punkte können die Label vollkommen beliebig sein.

Betrachte nun einen festen Lernalgorithmus und als Verteilung $\mathcal{D}$ die uniforme Verteilung auf T. Sei $h_{S,i}$ die Hypothese, die der Lernalgorithmus auf Sample S berechnet, wenn die Grundwahrheit f_i ist[1]. Wir möchten nun zeigen, dass

$$\max_i \mathbf{Pr}\left[\operatorname{err}_{\mathcal{D},f_i}(h_{S,i}) \geq \frac{1}{8}\right] \geq \frac{1}{7} \ .$$

Das heißt, dass es eine Grundwahrheit gibt, für die der Algorithmus schlecht ist. Definieren wir nun Zufallsvariablen Z_i (abhängig von S), so dass $Z_i = 1$ falls $\operatorname{err}_{\mathcal{D},f_i}(h_{S,i}) \geq \frac{1}{8}$, anderenfalls $Z_i = 0$.

In dieser Notation wollen wir zeigen, dass

$$\max_i \mathbf{Pr}\left[Z_i = 1\right] \geq \frac{1}{7} \ .$$

Hierfür ist es hinreichend, dass

$$\frac{1}{k}\sum_{i=1}^{k} \mathbf{Pr}\left[Z_i = 1\right] \geq \frac{1}{7} \ .$$

Da $\mathbf{Pr}\left[Z_i = 1\right] = \mathbf{E}\left[Z_i\right]$, ist diese Aussage mittels Linearität des Erwartungswertes äquivalent zu

$$\mathbf{E}\left[\sum_{i=1}^{k} Z_i\right] \geq \frac{k}{7} \ .$$

Betrachten wir ein festes $x \in T$, dann gibt es für jede Hypothese f_i genau eine Hypothese f_{-i}, die überall auf T mit f_i übereinstimmt, nur $f_i(x) \neq f_{-i}(x)$. Falls $x \notin S$, muss folglich gelten $h_{S,i} = h_{S,-i}$. Also muss entweder $h_{S,i}(x) \neq f_i(x)$ oder $h_{S,-i}(x) \neq f_{-i}(x)$ sein. Allgemeiner gesagt bedeutet dies, dass für alle $x \notin S$ gilt, dass $h_{S,i}(x) \neq f_i(x)$ für genau die Hälfte aller i.

Für jede feste Menge S mit $|S| \leq \frac{1}{2}|T|$ können wir also schreiben

$$\frac{1}{k}\sum_{i=1}^{k} \operatorname{err}_{\mathcal{D},f_i}(h_{S,i}) \geq \frac{1}{2}\frac{|T \setminus S|}{|T|} \geq \frac{1}{4} \ .$$

[1]Prinzipiell könnte $h_{S,i}$ auch randomisiert sein. Der Beweis würde genauso gelten. Der Einfachheit halber gehen wir aber davon aus, dass $h_{S,i}$ deterministisch von S und i abhängt.

Wenn wir S durch $\frac{d}{2}$ Züge aus $\mathcal{D}$ bestimmen, ist $|T \setminus S| \geq \frac{1}{2}|T|$.

Andererseits gilt auch

$$\sum_{i=1}^{k} \mathrm{err}_{\mathcal{D},f_i}(h_{S,i}) \leq \sum_{i=1}^{k} Z_i + \frac{1}{8}\left(k - \sum_{i=1}^{k} Z_i\right) = \frac{1}{8}k + \frac{7}{8}\sum_{i=1}^{k} Z_i \ ,$$

denn diejenigen i mit $Z_i = 1$ tragen höchstens 1, die übrigen höchstens $\frac{1}{8}$ zu der Summe bei.

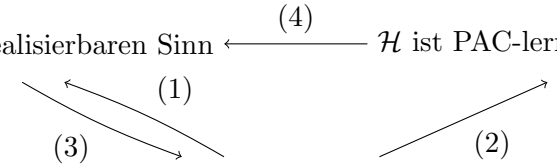
Folglich gilt also für jedes S immer

$$\sum_{i=1}^{k} Z_i \geq \frac{k}{7} \ .$$

Damit gilt die Ungleichung erst recht auch im Erwartungswert über S. □

3 Gesamtbild: PAC-Lernbarkeit

Zusammengenommen haben wir nun folgendes Bild von Implikationen.



Implikation (1) haben wir in den vergangenen Vorlesungen gezeigt. (2) folgt aus Satz 5.1, den wir nicht bewiesen haben. (3) ist die Aussage von Satz 5.5. (4) ist eine Übungsaufgabe. Insgesamt sind also alle drei Begriffe äquivalent.

Dies bedeutet übrigens nur, dass bei Hypothesenklassen mit endlicher VC-Dimension „genügend“ Samples für bei jeder Verteilung $\mathcal{D}$ ausreichen, um die beste Hypothese zu finden. Es bedeutet nicht, dass „genügend“ im realisierbaren und im agnostischen Fall gleich große Zahlen sind. Auch kann es bei Hypothesenklassen mit unendlicher VC-Dimension Verteilungen $\mathcal{D}$ geben, die Lernbarkeit ermöglichen.