AGML, Sommersemester 2020 Vorlesung 4 (5 Seiten)

Wachstumsfunktion und Agnostisches PAC-Lernen

Thomas Kesselheim Letzte Aktualisierung: 5. Mai 2020

1 Erinnerung: Wachstumsfunktion

Wir erinnern uns, dass eine Hypothesenklasse $\mathcal{H}$ eine Menge von Funktionen der Form $h\colon X \to \{-1,+1\}$ ist. Wir haben schon viele Beispiele gesehen, vor allem mit $X = \mathbb{R}$. Diese Hypothesenklassen enthalten jedoch nicht alle Funktionen sondern besitzen deutlich mehr Struktur. Dies wird formalisiert in der Wachstumsfunktion.

Definition 4.1. *Gegeben $S \subseteq X$, sei $\mathcal{H}|_S$ die Menge aller Hypothesen $h \in \mathcal{H}$ mit Definitionsbereich eingeschränkt auf S. Das heißt, $\mathcal{H}|_S = \{h|_S \mid h \in \mathcal{H}\}$.*

Die Wachstumsfunktion *von $\mathcal{H}$ ist definiert als $\Pi_{\mathcal{H}}(m) = \max_{S\subseteq X, |S|=m} |\mathcal{H}|_S|$.*

In der letzten Vorlesung haben wir ein extrem hilfreiches Werkzeug gesehen, um die Wachstumsfunktion zu beschränken: die VC-Dimension. Wir haben bewiesen, dass wenn die VC-Dimension d ist, auch $\Pi_{\mathcal{H}}(m) \leq \left(\frac{\mathrm{e}m}{\mathrm{d}}\right)^{\mathrm{d}}$ gilt. Das heißt, wenn die VC-Dimension endlich ist, wächst die Wachstumsfunktion nur polynomiell.

2 Subexponentielles Wachstum impliziert PAC-Lernbarkeit

Es steht noch der Beweis des Satzes aus der zweiten Vorlesung aus, dass derartiges subexponentielles Wachstum tatsächlich PAC-Lernbarkeit impliziert. Wir betrachten wieder eine Hypothesenklasse $\mathcal{H}$, eine Grundwahrheit $f \in \mathcal{H}$ und eine beliebige Wahrscheinlichkeitsverteilung $\mathcal{D}$ über X.

Satz 4.2. *Es seien $\epsilon > 0$ und $\delta > 0$ beliebig und*

$$m \geq \max\left\{\frac{8}{\epsilon}, \frac{2}{\epsilon}\log_2\left(\frac{2\Pi_{\mathcal{H}}(2m)}{\delta}\right)\right\} \ . \tag{1}$$

Betrachte ein Sample S von m Datenpunkten mit korrekten Labels gemäß f gezogen unabhängig und identisch verteilt aus $\mathcal{D}$. Es gilt mit Wahrscheinlichkeit mindestens $1-\delta$, dass alle $h \in \mathcal{H}$ mit $\mathrm{err}_S(h) = 0$ auch $\mathrm{err}_{\mathcal{D},f}(h) < \epsilon$ erfüllen.

Um Satz 4.2 zu zeigen, beweisen wir zunächst zwei Lemmata, die für sich genommen schon interessante Aussagen sind. Erst im Anschluss werden wir sie zum Beweis des Satzes zusammenfügen.

Wir halten zunächst fest, dass es eigentlich gar nicht mal sehr wahrscheinlich ist, dass eine feste Hypothese mit großem tatsächlichen Fehler auch „typischerweise“ einen großen Trainingsfehler hat.

Lemma 4.3. *Sei h eine Hypothese mit $\mathrm{err}_{\mathcal{D},f}(h) \geq \epsilon$ und sei S' eine Menge von m zufällig gezogenen Samples. Falls m Bedingung (1) erfüllt, dann gilt $\mathbf{Pr}\left[\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}\right] \geq \frac{1}{2}$.*

Beweis. Wir können uns das Zufallsexperiment vorstellen als m unabhängige Münzwürfe, wobei die Wahrscheinlichkeit für Kopf $p := \mathrm{err}_{\mathcal{D},f}(h) \geq \epsilon$ in jedem Wurf beträgt. Wir behaupten, dass wir mit Wahrscheinlichkeit mindestens $\frac{1}{2}$ mindestens $\frac{\epsilon}{2}m$ mal Kopf sehen.

Sei dazu Z die Anzahl Kopf in den Münzwürfen. Es gelten $\mathbf{E}[Z] = pm$ und $\mathrm{Var}[Z] = p(1-p)m$. Wegen $p \geq \epsilon$ gilt also nach der Tschebyschew-Ungleichung

$$\mathbf{Pr}\left[Z \leq \frac{\epsilon}{2}m\right] \leq \mathbf{Pr}\left[Z \leq \frac{p}{2}m\right] \leq \mathbf{Pr}\left[|Z - \mathbf{E}[Z]| \geq \frac{p}{2}m\right] \leq \frac{\mathrm{Var}[Z]}{\left(\frac{p}{2}m\right)^2} \leq \frac{p(1-p)m}{\left(\frac{p}{2}m\right)^2} = \frac{4(1-p)}{pm} \leq \frac{1}{2} \ ,$$

wobei wir im letzten Schritt $m \geq \frac{8}{\epsilon}$ und deshalb $pm \geq \epsilon m \geq 8$ benutzen. □

Die nächste Aussage ist, dass es, wenn *zwei Sample-Mengen* gezogen werden, eher unwahrscheinlich ist, dass es eine Hypothese gibt, die auf der einen Menge einen großen und auf der anderen Menge keinen Trainingsfehler hat.

Lemma 4.4. *Seien S und S' Mengen von m zufällig gezogenen Samples. Falls m Bedingung (1) erfüllt, dann gilt*

$$\mathbf{Pr}\left[\exists h' \in \mathcal{H} : \mathrm{err}_{S'}(h') \geq \frac{\epsilon}{2} \textit{ und } \mathrm{err}_S(h') = 0\right] \leq \frac{\delta}{2} \ .$$

Beweis. Wir beschreiben einen anderen aber äquivalenten Weg, um S und S' zu bestimmen: Wir ziehen $2m$ mal aus der Verteilung $\mathcal{D}$; sei das Ergebnis T. Jetzt ziehen wir m mal *ohne Zurücklegen* aus T und nennen das Ergebnis S. Schließlich ist S' der Rest aus T also $S' = T \setminus S$.

Betrachte nun eine feste Menge T und festes $h' \in \mathcal{H}$. Sei $h'(x) \neq f(x)$ für genau k Elemente aus T. Die einzige Art und Weise, wie $\mathrm{err}_{S'}(h') \geq \frac{\epsilon}{2}$ eintreten kann, ist dass $k \geq \frac{\epsilon}{2}m$.

Darüber hinaus ist die Wahrscheinlichkeit, dass h' keinen Fehler auf S macht gegeben als

$$\begin{aligned}\mathbf{Pr}\left[\mathrm{err}_S(h') = 0 \;\middle|\; T\right] &= \frac{2m-k}{2m} \cdot \frac{2m-k-1}{2m-1} \cdot \ldots \cdot \frac{m-k+1}{m+1} \\ &= \frac{m(m-1)\ldots(m-k+1)}{(2m)(2m-1)\ldots(2m-k+1)} \leq 2^{-k} \ .\end{aligned}$$

Hierbei gilt die zweite Gleichung, weil sich die alle Faktoren aus dem Zähler und dem Nenner kürzen bis auf die ersten k im Nenner und die letzten k im Zähler.

Das bedeutet, dass für festes h' und festes T

$$\mathbf{Pr}\left[\mathrm{err}_S(h') = 0 \text{ und } \mathrm{err}_{S'}(h') \geq \frac{\epsilon}{2} \;\middle|\; T\right] \leq \left\{ \begin{array}{ll} 0 & \text{falls } k < \frac{\epsilon}{2}m \\ 2^{-k} & \text{sonst} \end{array} \right\} \leq 2^{-\frac{\epsilon}{2}m} \ .$$

An dieser Stelle kommt die Wachstumsfunktion ins Spiel: die Menge T hat nur Größe $2m$. Das bedeutet, weil nur die Funktionswerte von h' auf T wichtig sind, dass es effektiv höchstens $\Pi_{\mathcal{H}}(2m)$ unterschiedliche Wahlen für h gibt. Deshalb gibt uns die Union Bound jetzt

$$\mathbf{Pr}\left[\exists h' \in \mathcal{H} : \mathrm{err}_S(h') = 0 \text{ und } \mathrm{err}_{S'}(h') \geq \frac{\epsilon}{2} \;\middle|\; T\right] \leq \Pi_{\mathcal{H}}(2m) 2^{-\frac{\epsilon}{2}m} \leq \frac{\delta}{2} \ .$$

Diese Schranke gilt für alle bedingten Wahrscheinlichkeiten, egal welche Menge T wir nutzen. Also gilt sie auch für die unbedingte Wahrscheinlichkeit. □

Beweis von Satz 4.2. Wir werden nun die Lemmata zusammenfügen. Sei A das Ereignis, dass es ein $h \in \mathcal{H}$ gibt mit $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$ aber $\mathrm{err}_S(h) = 0$. Wir möchten zeigen, dass $\mathbf{Pr}[A] \leq \delta$.

Um Lemma 4.4 anzuwenden, führen wir ein Hilfsereignis B ein. Sei dafür S' eine andere Menge von m Datenpunkten mit zugehörigen Labels, die auch unabhängig aus $\mathcal{D}$ gezogen sind. Sei B das Ereignis, dass es ein $h' \in \mathcal{H}$ gibt mit $\mathrm{err}_{S'}(h') \geq \frac{\epsilon}{2}$ aber $\mathrm{err}_S(h') = 0$. Gemäß Lemma 4.4 gilt $\mathbf{Pr}[B] \leq \frac{\delta}{2}$.

Darüber hinaus behaupten wir, dass $\mathbf{Pr}[B \mid A] \geq \frac{1}{2}$. Dafür sollten wir verstehen, was diese bedingte Wahrscheinlichkeit bedeutet. Ereignis A ist bereits eingetreten. Dieses hängt von der Menge S ab und sagt, dass es ein $h \in \mathcal{H}$ mit $\mathrm{err}_{\mathcal{D}}(h) \geq \epsilon$ aber $\mathrm{err}_S(h) = 0$. Damit Ereignis B eintritt, ist es nun hinreichend, dass $\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}$. (Es ist nicht gefordert, dass $h = h'$ ist, deshalb ist dies nur hinreichend aber nicht notwendig.) Nun können wir Lemma 4.3 nutzen. Die Wahrscheinlichkeit, dass für genau dieses h gilt $\mathrm{err}_{S'}(h) \geq \frac{\epsilon}{2}$ ist mindestens $\frac{1}{2}$.

Nun nutzen wir $\mathbf{Pr}[B] \geq \mathbf{Pr}[B \mid A]\,\mathbf{Pr}[A]$, um $\mathbf{Pr}[A] \leq \frac{\mathbf{Pr}[B]}{\mathbf{Pr}[B|A]}$ zu erhalten. Mit $\mathbf{Pr}[B] \leq \frac{\delta}{2}$ und $\mathbf{Pr}[B \mid A] \geq \frac{1}{2}$, folgt also $\mathbf{Pr}[A] \leq \delta$. □

3 Der Nicht-Realisierbare/Agnostische Fall

Bislang haben wir im Kontext von PAC-Learning nur den realisierbaren Fall behandelt. Das bedeutet, es gibt nicht nur eine Grundwahrheit $f\colon X \to \{-1, +1\}$, die die korrekten Labels angibt, sondern auch, dass f in der Hypothesenklasse $\mathcal{H}$ enthalten ist, die wir betrachten. Dies bedeutet insbesondere, dass es immer möglich ist, eine Hypothese zu finden, die keinen Trainingsfehler hat.

In typischen Fragen des Maschinellen Lernens ist diese Annahme jedoch eigentlich nie erfüllt. Die Merkmale beschreiben niemals die Wirklichkeit vollständig. Im Fall von Spam-Klassifikation mögen als Merkmale Worthäufigkeiten, IP-Adressen, Daten im Header und so weiter zur Verfügung stehen. Auf Basis dieser Information ist es aber unmöglich, alle E-Mails immer korrekt zu klassifizieren. Etwas philosophischer kann man sich auch fragen, ob es überhaupt eine klare Trennung zwischen Spam und erwünschten E-Mails gibt. Schließlich gibt es noch einen weiteren Aspekt: Selbst wenn es möglich wäre, eine Hypothesenklassen anzugeben, die eine perfekte Klassifikation ermöglichen würde, möchte man aus Effizienzgründen vielleicht eine weniger komplexe Klasse wählen.

Wie modellieren wir also Lernprobleme jenseits des realisierbaren Falls? Betrachten wir zunächst das linke Beispiel von Abbildung 1. Hier ist $X = [0, 1]^2$ und es gibt in der Tat eine Grundwahrheit $f\colon X \to \{-1, +1\}$, die allerdings relativ komplex ist. Nun könnte $\mathcal{H}$ die Menge aller linearen Klassifikatoren sein, das heißt, die Funktionen, die durch eine Gerade positive und negative Punkte trennen. In einem solchen Fall könnten wir weiterhin den tatsächlichen Fehler $\mathrm{err}_{\mathcal{D},f}(h)$ einer Hypothese h hinsichtlich einer Verteilung über Datenpunkte $\mathcal{D}$ und der Grundwahrheit f definieren als

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

Falls $f \notin \mathcal{H}$ ist, ist es nun aber nicht mehr möglich, dass $\mathrm{err}_{\mathcal{D},f}(h)$ beliebig klein wird.

Das rechte Beispiel ist komplexer. Hier gibt es keine Grundwahrheit. Es könnte beispielsweise sein, dass im Datenpunkte im grauen Bereich mit Wahrscheinlichkeit 50 % positiv und sonst negativ sind. Hierfür schauen wir uns Wahrscheinlichkeitsverteilungen über $X \times \{-1, +1\}$ an. Das heißt, diese Verteilung liefert einen zufälligen Datenpunkt mit Label. Äquivalent könnten wir auch wieder eine Verteilung über unbeschriftete Datenpunkte haben und dann für jeden von diesen eine Wahrscheinlichkeit eines positiven Labels.

Der tatsächlichen Fehler $\mathrm{err}_{\mathcal{D}}(h)$ einer Hypothese h hinsichtlich einer solchen Verteilung $\mathcal{D}$ über Datenpunkt-/Label-Paare ist definiert als

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

In beiden Fällen haben wir keine Hoffnung, eine Hypothese zu finden, sodass der tatsächliche Fehler beliebig klein wird. Stattdessen hoffen wir nun, möglichst nah an die bestmögliche Hypothese zu kommen.

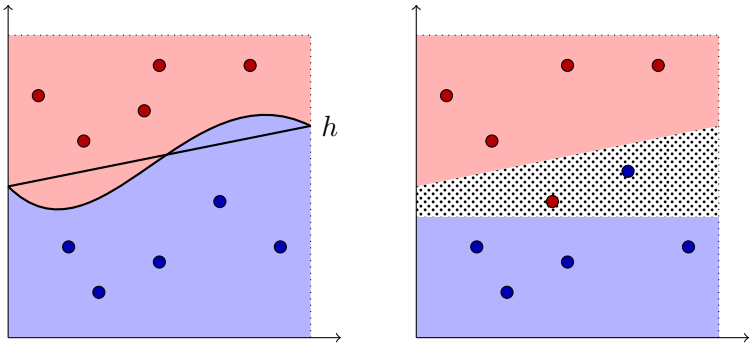

Abbildung 1: Beispiele von nicht-realisierbaren Fällen. Links gibt es keinen Hypothese h in unserer Klasse der linearen Separatoren, die mit der Grundwahrheit f auf allen Punkten übereinstimmt. Rechts sind im grauen Bereich die Labels zufällig; beispielsweise -1 oder $+1$ mit Wahrscheinlichkeit 50%. Es gibt also gar keine Funktion $f\colon X \to \{0,1\}$, die immer das korrekte Label zurückgibt.

Definition 4.5. *Eine Hypothesenklasse $\mathcal{H}$ ist PAC-lernbar (im agnostischen Sinn), wenn es eine Funktion $m_{\mathcal{H}}$ und einen Lernalgorithmus gibt, der für alle $\epsilon, \delta > 0$ und jede Verteilung $\mathcal{D}$ über Datenpunkt-/Label-Paare mithilfe eines zufälligen Samples S der Größe mindestens $m_{\mathcal{H}}(\epsilon, \delta)$ aus $\mathcal{D}$ gezogen, eine Hypothese $h_S \in \mathcal{H}$ berechnet, sodass*

$$\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D}}(h_S) < \min_{h' \in \mathcal{H}} \mathrm{err}_{\mathcal{D}}(h') + \epsilon\right] \geq 1 - \delta \ .$$

Agnostisch bezieht sich hierbei darauf, dass nicht bekannt, aber auch unerheblich ist, ob es eine Grundwahrheit (in $\mathcal{H}$ bzw. allgemein) gibt, oder nicht.