AGML, Sommersemester 2020 Vorlesung 2 (3 Seiten)

Wachstumsfunktion

Thomas Kesselheim Letzte Aktualisierung: 24. April 2020

1 Wiederholung: PAC-lernbar (Realisierbarer Fall)

Unsere Aufgabe ist es, Datenpunkte aus einer Menge X zu klassifizieren, beispielsweise $X \subseteq \mathbb{R}$. Die Label werden binär sein, das heißt -1 oder 1. Beispielsweise könnte X die Menge aller E-Mails sein und die Labels habe die Bedeutung „nicht Spam“ oder „Spam“. Unser Ziel ist es, dass wir für jeden Datenpunkt x, den wir als Eingabe erhalten, das korrekte Label $y \in \{-1, 1\}$ vorhersagen zu können.

Es gibt eine Klasse von Hypothesen $\mathcal{H}$. Jede hat die Form $h\colon X \to \{-1, 1\}$. Wir nehmen an, dass wir im *realisierbaren Fall* sind. Das heißt, es gibt eine Grundwahrheit $f \in \mathcal{H}$, die eine unserer möglichen Hypothesen ist, und das korrekte Label für $x \in X$ ist immer $f(x)$. Wir möchte nun eine Funktion $h \in \mathcal{H}$ finden, die möglichst ähnlich zum korrekten f ist. Dafür steht uns aber nur eine begrenzte Anzahl Samples mit korrekten Labels zur Verfügung.

Wir erinnern uns an die Definition von PAC-Lernbarkeit.

Definition 2.1. *Eine Hypothesenklasse $\mathcal{H}$ heißt* PAC-lernbar *(im realisierbaren Sinn), wenn es eine Funktion $m_{\mathcal{H}}$ und einen Lernalgorithmus $\mathcal{A}$ gibt, sodass der Algorithmus für alle $\epsilon, \delta > 0$, jede Verteilung $\mathcal{D}$ und alle $f \in \mathcal{H}$, gegeben ein Sample S von Größe mindestens $m_{\mathcal{H}}(\epsilon, \delta)$ von Datenpunkten mit korrekten Labels, eine Hypothese $h_S \in \mathcal{H}$ berechnet, sodass $\mathbf{Pr}\left[\mathrm{err}_{\mathcal{D},f}(h_S) < \epsilon\right] \geq 1 - \delta$.*

Hierbei ist $\mathrm{err}_{\mathcal{D},f}(h) := \mathbf{Pr}_{x\sim\mathcal{D}}\left[h(x) \neq f(x)\right]$ der tatsächliche Fehler von h. Zwei Beispiele dafür haben wir bereits gesehen. Heute wollen wir uns das Thema etwas allgemeiner anschauen.

2 Minimierung des Trainingsfehlers

Wir werden uns allgemeiner Algorithmen anschauen, die den Trainingsfehler minimieren.

Definition 2.2. *Der* Trainingsfehler *(oder* empirisches Risiko*)* $\mathrm{err}_S(h)$ *einer Hypothese h hinsichtlich einer Trainingsmenge S ist*

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

Im realisierbaren Fall gilt für die Grundwahrheit f immer $\mathrm{err}_S(f) = 0$ für alle S. Unsere Algorithmen aus der letzten Vorlesung berechneten jedoch auch jeweils Hypothesen h, sodass $\mathrm{err}_S(h) = 0$. Auch diese minimieren also den Trainingsfehler. Unsere Frage heute wird sein, den tatsächlichen Fehler von Hypothesen zu beschränken, die den Trainingsfehler minimieren.

3 Endliche Hypothesenklassen

Wir betrachten zunächst den einfachen Fall, dass die Menge $\mathcal{H}$ endlich ist, wenn auch ansonsten beliebig.

Satz 2.3. *Wenn $m \geq \frac{1}{\epsilon}\ln\left(\frac{|\mathcal{H}|}{\delta}\right)$, dann 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.*

Beweis. Wir betrachten zunächst ein festes $h \in \mathcal{H}$ mit $\mathrm{err}_{\mathcal{D},f}(h) \geq \epsilon$, das heißt, der tatsächliche Fehler von h ist mindestens ϵ. Nun gilt

$$\begin{aligned}\mathbf{Pr}\left[\mathrm{err}_S(h) = 0\right] &= \mathbf{Pr}\left[h(x_1) = y_1, \ldots, h(x_m) = y_m\right] \\ &= \mathbf{Pr}\left[h(x_1) = y_1\right] \cdot \ldots \cdot \mathbf{Pr}\left[h(x_m) = y_m\right] \leq (1-\epsilon)^m \leq \mathrm{e}^{-\epsilon m} \ .\end{aligned}$$

Das heißt, dass die Wahrscheinlichkeit, dass h keinen Trainingsfehler hat, höchstens $\mathrm{e}^{-\epsilon m}$ ist.

Um die Gesamtwahrscheinlichkeit zu beschränken, dass es irgendeine Hypothese gibt, die zwar keinen Trainingsfehler, aber großen tatsächlichen Fehler hat, benutzen wir die sogenannte Union Bound.

Lemma 2.4 (Union Bound)**.** *Es seien $\mathcal{E}_1, \ldots, \mathcal{E}_n$ (nicht notwendigerweise disjunkte) Ereignisse. Dann gilt*

$$\mathbf{Pr}\left[\bigcup_{i=1}^{n} \mathcal{E}_i\right] \leq \sum_{i=1}^{n} \mathbf{Pr}\left[\mathcal{E}_i\right] \ .$$

Der Beweis der Union Bound folgt durch induktive Anwendung von $\mathbf{Pr}\left[A \cup B\right] = \mathbf{Pr}\left[A\right] + \mathbf{Pr}\left[B\right] - \mathbf{Pr}\left[A \cap B\right] \leq \mathbf{Pr}\left[A\right] + \mathbf{Pr}\left[B\right]$.

Um nun die Union Bound anzuwenden, definieren wir für jede Hypothese $h \in \mathcal{H}$ das Ereignis $\mathcal{E}_h$, dass $\mathrm{err}_S(h) = 0$.

Nun gilt

$$\begin{aligned}\mathbf{Pr}\left[\exists h \in \mathcal{H} : \mathrm{err}_{\mathcal{D},f}(h) \geq \epsilon \text{ und } \mathrm{err}_S(h) = 0\right] &= \mathbf{Pr}\left[\bigcup_{h \in \mathcal{H}, \mathrm{err}_{\mathcal{D},f}(h) \geq \epsilon} \mathcal{E}_h\right] \\ &\leq \sum_{h \in \mathcal{H}, \mathrm{err}_{\mathcal{D},f}(h) \geq \epsilon} \mathbf{Pr}\left[\mathrm{err}_S(h) = 0\right] \\ &\leq |\mathcal{H}|\mathrm{e}^{-\epsilon m} \leq \delta \ . \qquad \square\end{aligned}$$

4 Wachstumsfunktion

Dieses Ergebnis nützt uns natürlich nichts, wenn $\mathcal{H}$ unendlich ist. Wir haben allerdings schon Beispiele gesehen, dass auch unendliche Hypothesenklassen PAC-lernbar sein können, beispielsweise die Schwellenwertfunktionen. Diese haben eine Struktur, die wir ausnutzen können. Dies können wir wie folgt formalisieren.

Definition 2.5. *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 $\Pi_{\mathcal{H}}(m) = \max_{S \subseteq X, |S|=m} |\mathcal{H}|_S|$.*

Weil die Abbildungen in $\mathcal{H}|_S$ von S nach $\{-1, +1\}$ abbilden, können es nicht mehr als 2^m verschiedene sein, weil es nicht mehr Abbildungen gibt. Somit muss immer $\Pi_{\mathcal{H}}(m) \leq 2^m$ gelten. Häufig sind die Werte von $\Pi_{\mathcal{H}}$ jedoch viel kleiner.

Beispiel 2.6. *Betrachte $X = \mathbb{R}$ und $\mathcal{H}$ als die Klasse der Schwellenwertfunktionen*

$$h_{a'}(x) = \begin{cases} +1 & \textit{falls } x \geq a' \\ -1 & \textit{sonst} \end{cases}$$

Für $S = \{2, 3, 4\}$ besteht $\mathcal{H}|_S$ aus folgenden vier Funktionen:

$$x \mapsto -1 \quad \textit{für alle } x \qquad\qquad x \mapsto +1 \quad \textit{für alle } x$$

$$x \mapsto \begin{cases} -1 & \textit{für } x = 2 \textit{ oder } x = 3 \\ +1 & \textit{für } x = 4 \end{cases} \qquad x \mapsto \begin{cases} -1 & \textit{für } x = 2 \\ +1 & \textit{für } x = 3 \textit{ oder } x = 4 \end{cases}$$

Es gibt noch vier weitere Funktionen $\{2, 3, 4\} \to \{-1, +1\}$. Diese lassen sich aber nicht über einen Schwellenwert realisieren.

Allgemein gilt $\Pi_{\mathcal{H}}(m) = m + 1$, denn es gibt nur $m + 1$ mögliche „Umschaltpunkte“ von -1 auf $+1$. Das heißt, die Funktion wächst deutlich schwächer als 2^m.

Der folgende Satz zeigt, dass wir in der Aussage von Satz 2.3 im Wesentlichen die Größe von $\mathcal{H}$ durch die Wachstumsfunktion ersetzen können.

Satz 2.7. *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.

Bevor wir mit dem Beweis dieses Satzes beginnen, schauen wir uns zunächst die Aussage etwas genauer an. Sie hat grundsätzlich die Struktur der Aussage, wie wir sie für PAC-Lernbarkeit brauchen. Wenn m Bedingung (1) erfüllt, dann führt beliebiger Lernalgorithmus, der den Trainingsfehler minimiert, zu einem tatsächlichen Fehler von höchstens ϵ mit Wahrscheinlichkeit mindestens $1 - \delta$.

Wann gilt jedoch Bedingung 1 und wann ist sie überhaupt für alle ϵ und δ erfüllbar? Schauen wir uns nur noch $m \geq \frac{8}{\epsilon}$ an, dann brauchen wir noch

$$m \geq \frac{2}{\epsilon}\log_2\left(\frac{2\Pi_{\mathcal{H}}(2m)}{\delta}\right) = \frac{2}{\epsilon}\log_2\left(\Pi_{\mathcal{H}}(2m)\right) + \frac{2}{\epsilon}\log_2\left(\frac{2}{\delta}\right) \quad \Leftrightarrow \quad \frac{m - \log_2\left(\frac{2}{\delta}\right)}{\log_2\left(\Pi_{\mathcal{H}}(2m)\right)} \geq \frac{2}{\epsilon} \ .$$

Wenn $\Pi_{\mathcal{H}}(2m) = 2^{2m}$ (die triviale Schranke), dann ist $\log_2\left(\Pi_{\mathcal{H}}(2m)\right) = 2m$. Die Ungleichung ist also für sinnvolle ϵ (d.h. $\epsilon < 1$) nicht erfüllbar.

Wächst hingegen $\log_2\left(\Pi_{\mathcal{H}}(2m)\right)$ schwächer als m, das heißt, $\log_2\left(\Pi_{\mathcal{H}}(2m)\right) = o(m)$, dann muss m nur ausreichend groß genug gewählt werden, um die Schranke zu erfüllen.

Im Beispiel mit den Schwellenwertfunktionen ist dies der Fall. Es gilt $\Pi_{\mathcal{H}}(2m) = 2m + 1$. Nun gilt also für alle $\delta > 0$, dass

$$\frac{m - \log_2\left(\frac{2}{\delta}\right)}{\log_2\left(\Pi_{\mathcal{H}}(2m)\right)} = \frac{m - \log_2\left(\frac{2}{\delta}\right)}{\log_2(2m + 1)} \to \infty \quad \text{für } m \to \infty \ .$$

Egal, wie ϵ und δ als gewählt sind, für genügend große m ist Bedingung 1 immer erfüllt.