Die Wachstumsfunktion $m_H(N)$

M. Gruber

KW 51

Wir kennen die Hoeffdingungleichung für einen endlichen Hypothesenraum $H = \{h_1, \ldots, h_M\}$,

$$\mathbf{P}(|\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{out}}(h)| > \varepsilon) \leq 2Me^{-2\varepsilon^2 N}. \tag{1}$$

N steht für die Mächtigkeit der Trainingsdaten, M für die des Hypothesenraums und ε für die Toleranz. Die Schwäche der Ungleichung (1) ist das Wachstum der rechten Seite in Abhängigkeit von M. Schreibt man für die rechte Seite δ und löst nach ε auf, bekommt man

$$\varepsilon = \sqrt{\frac{1}{2N} \ln \frac{2M}{\delta}} \tag{2}$$

und sieht, wie sich ein wachsendes M auf die Toleranz ε auswirkt. Beim PLA ist $|M| = \infty$ und (1) hilft gar nicht.

Wir werden sehen, dass man M durch eine Wachstumsfunktion $m_H(N)$ ersetzen kann. Diese wächst in günstigen Fällen polynomial in N. Dann dämpft $e^{-2\varepsilon^2 N}$ stärker als $m_H(N)$ wächst und die Aussichten auf erfolgreiches Lernen durch einen Lernalgorithmus nehmen mit N zu.

Sei H ein Hypothesenraum beliebiger Größe. Für jedes $h \in H$ gelte $h : X \longrightarrow \{-1, 1\}$. Auf einer N-elementigen Trainingsmenge $\{x_1, \ldots, x_N\}$ stellt sich eine Hypothese h durch die Dichotomie $[h(x_1), \ldots, h(x_n)]$ dar, d.h. durch einen N-stelligen Vektor mit Komponenten -1 oder 1. Hypothesen, die dieselbe Dichotomie erzeugen, sind für uns äquivalent. Die Äquivalenzrelation zerlegt den Hypothesenraum in endlich viele Äquivalenzklassen. Das ist der wesentliche Punkt.

Definition 1 (Wachstumsfunktion) *Sei $X = \{x_1, \ldots, x_n\}$ eine Lernmenge, H ein Hypothesenraum mit $h : X \longrightarrow \{-1, 1\}$ für $h \in H$, $H(x_1, \ldots, x_n)$ die Menge der von H erzeugten Dichotomien und $|H(x_1, \ldots, x_n)|$ die Mächtigkeit (Anzahl der Elemente) der Dichotomienmenge. Die Funktion*

$$m_H(N) = \max_{\{x_1, \ldots, x_N\} \subset X} |H(x_1, \ldots, x_N)| \tag{3}$$

ist die Wachstumsfunktion zu X und H.

Die Wachstumsfunktion gibt die größtmögliche Anzahl von Dichotomien an, die H auf einer N-elementigen Teilmenge von X erzeugen kann. Sie ist zugleich die größtmögliche Anzahl von Äqui-

valenzklassen (im Sinne der oben eingeführten Äquivalenzrelation), die durch eine N-elementige Teilmenge von X auf der Hypothesenmenge erzeugt werden kann.

Bemerkung 1 *Es ist stets* $m_H(N) \leq 2^N$.

Von Interesse sind die Fälle, in denen $m_H(N)$ das exponentielle Wachstum unterschreitet.

Definition 2 (Breakpoint) *Seien* X, H *gegeben. Ein* $k \in \mathbf{N}$ *mit* $m_H(k) < 2^k$ *heißt Breakpoint.*

Beispiel 1 *Beim PLA ist im zweidimensionalen Fall* $k = 4$ *ein Breakpoint:* $m_H(4) = 14 < 2^4$.

Definition 3 *Sei* $B(N, k)$ *die größtmögliche Zahl von Dichotomien auf einer* N*-elementigen Trainingsmenge, unabhängig von* H*, aber unter der Bedingung, dass* k *ein Breakpoint ist:*

$$B(N, k) = \max_{H \text{ with Breakpoint } k} m_H(N).$$

Über die abstrakte Größe $B(N, k)$ kann man erstaunlich konkrete Aussagen machen.

Satz 1 *Es ist*

$$B(N,k) \begin{cases} = 1 & \text{für} \quad N \geq 1 \text{ und } k = 1, \\ = 2 & \text{für} \quad N = 1 \text{ und } k > 1, \\ \leq B(N-1,k) + B(N-1,k-1) & \text{für} \quad N \geq 2 \text{ und } k \geq 2. \end{cases}$$

Beweis Wenn $k = 1$ ein Breakpoint ist, können zulässige H's maximal eine Dichotomie auf einer einelementigen Lernmenge erzeugen. Solche H's können aber auch auf einer N-elementigen Lernmenge nicht mehr als eine Dichotomie erzeugen. Sonst gäbe es ein x und Hypothesen h_1, h_2 mit $h_1(x) \neq h_2(x)$ und $k = 1$ wäre doch kein Breakpoint. Also ist $B(N, 1) = 1$.

Wenn die Lernmenge einelementig ist ($N = 1$), kann sie keine Teilmenge mit mehr als einem Element ($k > 1$) enthalten. Die Nebenbedingung "k Breakpoint" stellt in diesem Fall keine Einschränkung dar. Ohne Einschränkung ist $m_H(1) = 2$. Also ist $B(1, k) = 2$ für $k > 1$.

Nun zum Fall $N \geq 2, k \geq 2$. Sei $S = H(x_1, \ldots, x_N)$ die Dichotomienmenge einer Realisierung des Maximums $B(N, k)$ mit konkreten X und H. S lässt sich in die disjunkten Mengen S_1, S_2^+, S_2^- zerlegen, wobei diese folgendermaßen charakterisiert sind:

$$\begin{aligned} S_1 &= \{d = [d_1, \ldots, d_N] \in S \mid [d_1, \ldots, d_{N-1}] \text{ kommt nur einmal in } S \text{ vor}\} \\ S_2^+ &= \{d = [d_1, \ldots, d_N] \in S \mid [d_1, \ldots, d_{N-1}] \text{ kommt doppelt in } S \text{ vor und } d_N = 1\} \\ S_2^- &= \{d = [d_1, \ldots, d_N] \in S \mid [d_1, \ldots, d_{N-1}] \text{ kommt doppelt in } S \text{ vor und } d_N = -1\} \end{aligned}$$

Sei $B(N, k) = \alpha + 2\beta$ mit $|S_1| = \alpha$, $|S_2^+| = \beta$ und $|S_2^-| = \beta$.

Streicht man aus den Dichotomien von $S_1 \cup S_2^+$ die letzte Komponente, bekommt man $\alpha + \beta$ verschiedenen Dichotomien $[h(x_1), \ldots, h(x_{N-1})]$, wobei die $x_1, \ldots, x_{N-1}$ einer Lernmenge X entstammen, für die (unter H) k ein Breakpoint ist. Es muss also $\alpha + \beta \leq B(N-1, k)$ sein.

Betrachte $H(x_1, \ldots, X_{N-1})$. Es gibt keine Teilmenge $\{x_{i_1}, \ldots, x_{i_{k-1}}\}$ von $\{x_1, \ldots, x_{N-1}\}$ mit der Eigenschaft $|H(x_{i_1}, \ldots x_{i_{k-1}})| = 2^{k-1}$, sonst hätte man $|H(x_{i_1}, \ldots x_{i_{k-1}}, x_N)| = 2^k$ und k wäre kein Breakpoint von X und H. $H(x_1, \ldots, X_{N-1})$ repräsentiert also ein Lernmodell mit $N-1$ Elementen und Breakpoint $k-1$. Damit ist $\beta = |H(x_1, \ldots, X_{N-1})| \leq B(N-1, k-1)$.

Insgesamt ist also $B(N,k) = \alpha + 2\beta \leq B(N-1,k) + B(N-1,k-1)$ für $N \geq 2, k \geq 2$. □

Bemerkung 2 *Die Funktion*

$$F(N,k) \begin{cases} = 1 & \text{für} \quad N \geq 1 \text{ und } k = 1, \\ = 2 & \text{für} \quad N = 1 \text{ und } k > 1, \\ = F(N-1,k) + F(N-1,k-1) & \text{für} \quad N \geq 2 \text{ und } k \geq 2. \end{cases}$$

dominiert B. Hier ist ein Tableau mit den ersten Werten von F:

	k=1	2	3	4	5
N=1	1	2	2	2	2
2	1	3	4	4	4
3	1	4	7	8	8
4	1	5	11	15	16
5	1	6	16	26	31

Satz 2 (Sauer's Lemma) *[1], p.48* $B(N,k) \leq \sum_{0 \leq i \leq k-1} \binom{N}{i}$.

Bemerkung 3 $\sum_{0 \leq i \leq k-1} \binom{N}{i} = O(N^{k-1})$, *denn* $\sum_{0 \leq i \leq k-1} \binom{N}{i} = \binom{N}{0} + \binom{N}{1} + \ldots + \binom{N}{k-1}$.

Bemerkung 4 *Ist k ein Breakpoint, so ist* $m_H(N) = O(N^{k-1})$.

Beispiel 2 (Positiver Strahl) *Sei* $X = \mathbf{R}$ *und* $H = \{h : \mathbf{R} \longrightarrow \{-1, 1\} \mid h(x) = \operatorname{sgn}(x - a)\}$. *Ein Breakpoint ist* $k = 2$. *Es ist* $m_H(N) = N + 1$.

Beispiel 3 (Positives Intervall) *Sei* $X = \mathbf{R}$ *und* $H = \{h_I \mid I \text{ Intervall}, h_I(x) = 2[x \in I] - 1\}$. *Ein Breakpoint ist* $k = 3$. *Es ist* $m_H(N) = \binom{N+1}{2}$.

Referenzen

[1] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin. *Learning From Data.* AMLBook, 2012.