Generalisierungstheorie

M. Gruber

KW 51

Satz 1. [Hoeffding-Ungleichung] *Seien $V_i, i = 1, \ldots, N$ unabhängige Zufallsvariablen mit $\mathbf{P}(V_i \in [0,1]) = 1$ und sei $\bar{V} = \frac{1}{N}\sum_{1\leq i\leq N} V_i$. Dann gilt*

$$\mathbf{P}(\bar{V} - \mathbf{E}\,\bar{V} \geq \varepsilon) \leq e^{-2\varepsilon^2 N} \quad \textit{und} \quad \mathbf{P}(|\bar{V} - \mathbf{E}\,\bar{V}| \geq \varepsilon) \leq 2e^{-2\varepsilon^2 N}. \tag{1}$$

Bemerkung 1. *Aus (1) folgt für eine Hypothese $h \in H$*

$$\mathbf{P}(\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{out}}(h) \geq \varepsilon) \leq e^{-2\varepsilon^2 N} \tag{2}$$

$$\mathbf{P}(|\,\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{out}}(h)| \geq \varepsilon) \leq 2e^{-2\varepsilon^2 N}. \tag{3}$$

Bemerkung 2. *Aus (2) bzw. (3) folgt für $H = \{h_1, \ldots, h_M\}$*

$$\mathbf{P}(\sup_{h\in H}(\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{out}}(h)) \geq \varepsilon) \leq Me^{-2\varepsilon^2 N} \tag{4}$$

$$\mathbf{P}(\sup_{h\in H}|\,\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{out}}(h)| \geq \varepsilon) \leq 2Me^{-2\varepsilon^2 N}. \tag{5}$$

Definition 1. [Wachstumsfunktion] *Die Hypothesenmenge H bestehe aus Hypothesen $h : \mathcal{X} \to \{-1, 1\}$. Für Lernmengen $\{x^{(1)}, \ldots, x^{(N)}\} \subset \mathcal{X}$ sei $H(x^{(1)}, \ldots, x^{(N)}) = \{[h(x^{(1)}), \ldots, h(x^{(N)})] \mid h \in H\}$ die Menge der von H auf $\{x^{(1)}, \ldots, x^{(N)}\}$ erzeugten Dichotomien. Sei $|H(x_1, \ldots, x_n)|$ die Anzahl der Elemente von $H(x^{(1)}, \ldots, x^{(N)})$. Die Funktion*

$$m_H(N) = \max_{\{x_1,\ldots,x_N\}\subset\mathcal{X}} |H(x_1, \ldots, x_N)| \tag{6}$$

ist die Wachstumsfunktion zu X und H.

Bemerkung 3. *Es ist stets $m_H(N) \leq 2^N$.*

Definition 2. [Breakpoint] *Seien $\mathcal{X}$ und 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:*

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

Satz 2. *Es ist*

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

Bemerkung 4. *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 3. [Sauer's Lemma] *[1], p.48)* $B(N,k) \leq \sum_{0 \leq i \leq k-1} \binom{N}{i}$.

Bemerkung 5. $\sum_{0\leq i\leq k-1}\binom{N}{i} = O(N^{k-1})$, *denn* $\sum_{0\leq i\leq k-1}\binom{N}{i} = \binom{N}{0} + \binom{N}{1} + \ldots + \binom{N}{k-1}$.

Bemerkung 6. *Ist k ein Breakpoint, so ist $m_H(N) = O(N^{k-1})$.*

Beispiel 2. [Positiver Strahl] *Sei $\mathcal{X} = \mathbf{R}$ 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 $\mathcal{X} = \mathbf{R}$ und*

$$H = \{h_I \mid I \text{ Intervall}, h_I(x) = 2[x \in I] - 1\}.$$

Ein Breakpoint ist $k = 3$. Es ist $m_H(N) = \binom{N+1}{2}$.

Satz 4. [Vapnik und Chervonenkis, 1971] *Sind Inputmenge $\mathcal{X}$ und Hypothesenmenge H gegeben, so gilt, wenn die Lernmenge N Elemente umfasst,*

$$\mathbf{P}\Big(\sup_{h\in H}\big|\mathrm{Err}_{\mathrm{in}}(h)-\mathrm{Err}_{\mathrm{out}}(h)\big|>\epsilon\Big)\leq 4m_H(2N)e^{-\frac{1}{8}\varepsilon^2 N}. \tag{7}$$

Die Vapnik-Chervonenkis-Ungleichung gilt als das wichtigste theoretische Ergebnis für statistisches Lernen.

Definition 4. [VC-Dimension] *Seien Inputmenge $\mathcal{X}$ und Hypothesenmenge H gegeben. Die größte Zahl N mit $m_H(N) = 2^N$ nennt man die VC-Dimension von H. Man bezeichnte sie mit $d_{VC}(H)$.*

$$d_{VC}(H) = \max\{N \mid m_H(N) = 2^N\}. \tag{8}$$

Ist dieses Maximum nicht definiert, setzt man $d_{VC}(H) = \infty$. Wenn der Zusammenhang klar ist, schreibt man kurz d_{VC} für $d_{VC}(H)$.

Bemerkung 7. *Ist $d_{VC} < \infty$, so ist jedes $k > d_{VC}$ ein Breakpoint.*

Bemerkung 8. *Ist $d_{VC} < \infty$, so ist $m_H(N) \leq \sum_{0 \leq i \leq d_{VC}} \binom{N}{i} = O(N^{d_{VC}})$.*

Bemerkung 9. *In der Lernsituation "Positiver Strahl" ist $d_{VC} = 1$. Beim 2D-Perzeptron ist $d_{VC} = 3$. In der Lernsituation "Konvexe Mengen" ist $d_{VC} = \infty$.*

Bemerkung 10. *Ist $d_{VC}(H) < \infty$, so ist jedes $h \in H$ im Allgemeinfall ebenso "gut" wie auf der Trainingsmenge, unabhängig vom Lernalgorithmus, von der Wahrscheinlichkeitsverteilung der Daten und von der Zielfunktion.*

Satz 5. *Beim Perzeptron-Modell der Dimension d ist $d_{VC} = d + 1$.*

Beweis Im Fall $d = 2$ haben wir $d_{VC} = 3$ schon gesehen.

Den allgemeinen Fall zeigen wir in zwei Schritten. Erst zeigen wir $d_{VC} \geq d+1$, dann $d_{VC} \leq d+1$.

- Betrachte die $(d + 1) \times (d + 1)$-Datenmatrix

$$X = \begin{bmatrix} x_1 & x_2 & \cdots & x_{d+1} \end{bmatrix}^T = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 1 & 1 & \cdots & 1 \end{bmatrix}.$$

  Sie hat maximalen Rang und ist invertierbar. Mit X ist jedes Gleichungssystem $Xw = y$ lösbar. Insbesondere ist $\mathrm{sgn}(Xw) = y$ für jeden Label-Vektor y lösbar. Für den Hypothesenraum H der Gewichtsvektoren w heißt dies, dass er auf der Lernmenge $\{x_1, \ldots, x_{d+1}\}$ alle 2^{d+1} verschiedenen Label-Vektoren erzeugen kann. Also ist $d_{VC} \geq d + 1$.

- Nun zeigen wir: eine $(d+2) \times (d+1)$-Datenmatrix $X = \begin{bmatrix} x_1 & x_2 & \cdots & x_{d+2} \end{bmatrix}^T$ kann nicht alle 2^{d+2} Label-Vektoren der Länge $d+2$ erzeugen. Eine solche Datenmatrix hat einen Zeilenrang $\leq d+1$. Es gibt ein $a \in \mathbf{R}^{d+2}$ mit $a \neq 0$ und $a^T X = 0$. O.B.d.A. sei $a_j = -1$ für ein bestimmtes j. Dann ist $x_j^T = \sum_{i \neq j} a_i x_i^T$. Für einen beliebigen Gewichtsvektor w aus dem Hypothesenraum ist

$$x_j^T w = \sum_{i \neq j} a_i x_i^T w. \tag{9}$$

  Man sieht nun, dass eine Dichotomie y mit $y_j = -1$ und $y_i = \operatorname{sgn}(a_i)$ für $i \neq j$ nicht realisierbar ist. Denn in (9) wäre sonst die linke Seite negativ, die rechte aber positiv. Also ist $d_{VC} \leq d+1$.

□

Satz 6. *[2] Consider a two-layer feedforward network $\mathcal{N}$ with input domain $\mathcal{X} = \{-k, \ldots, k\}^n$ (for $k \in \mathbf{N}$) and first-layer computation gates with the standard sigmoid activation function (the output gate being a linear threshold gate). Let w be the total number of parameters in the network. Then $d_{VC}(\mathcal{N}) = O(w \log(wk))$.*

Referenzen

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

[2] P. L. Bartlett and W. Maass. Vapnik-Chervonenkis dimension of neural nets. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1188–1192. MIT Press, 2003. Second Edition.