Generalisierungstheorie

M. Gruber

KW 51

1 Vapnik-Chervonenkis-Ungleichung

Wir folgen [1] pp.187–192.

Satz 1 (Vapnik und Chervonenkis, 1971) *Sind Inputmenge X und Hypothesenmenge H gegeben, so gilt, wenn die Lernmenge N Elemente umfasst,*

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

Die Vapnik-Chervonenkis-Ungleichung gilt als das wichtigste theoretische Ergebnis für statistisches Lernen. Der Beweis stützt sich auf eine Reihe von Hilfssätzen.

Satz 2 *([1], Lemma A.2.) Angenommen, es liegen zwei unabhängige N-elementige Lernmengen $D, D'\subset X$ vor. Für eine Hypothese $h\in H$ bezeichne $\mathrm{Err}_{\mathrm{in}}(h)$ den in-sample error auf D und $\mathrm{Err}_{\mathrm{in}}{}'(h)$ den in-sample error auf D'. Für die Ereignisse*

$$A=\big\{sup_{h\in H}|\mathrm{Err}_{\mathrm{in}}(h)-\mathrm{Err}_{\mathrm{in}}{}'(h)|>\tfrac{\varepsilon}{2}\big\}\quad \textit{und}\quad B=\{sup_{h\in H}|\mathrm{Err}_{\mathrm{in}}(h)-\mathrm{Err}_{\mathrm{out}}(h)|>\varepsilon\}$$

gilt dann

$$(1-2e^{-(1/2)\varepsilon^2 N})\,\mathbf{P}(B)\leq\mathbf{P}(A). \tag{2}$$

Beweis Die Ereignisse A und B liegen in $\Omega=X^{2N}$. Das Ereignis A hängt von allen Komponenten eines $[D,D']\in\Omega$ ab, das Ereignis B nur von den ersten N Komponenten. Wahrscheinlichkeiten werden mit dem Produktmaß $\mathbf{P}^{2N}$ gemessen. Nur der Einfachheit halber schreiben wir dieses als $\mathbf{P}$.

Im nichttrivialen Fall $\mathbf{P}(B)>0$ gilt immer $\mathbf{P}(A)\geq\mathbf{P}(A\cap B)=\mathbf{P}(A\mid B)\,\mathbf{P}(B)$. Wir werden gleich sehen, dass in unserem Fall $\mathbf{P}(A\mid B)\geq(1-2e^{-(1/2)\varepsilon^2 N})$ ist, falls $\mathbf{P}(B)>0$ ist.

Für jedes D, das in einem $[D,D']\in B$ vorkommt, wählen wir eine Hypothese $h^{(D)}$ mit $|\mathrm{Err}_{\mathrm{in}}(h^{(D)})-\mathrm{Err}_{\mathrm{out}}(h^{(D)})|>\varepsilon$. Wegen $A\supset\{|\mathrm{Err}_{\mathrm{in}}(h^{(D)})-\mathrm{Err}_{\mathrm{in}}{}'(h^{(D)})|>\frac{\varepsilon}{2}\}$ gilt

$$\mathbf{P}(A\mid B)\geq\mathbf{P}(\{|\mathrm{Err}_{\mathrm{in}}(h^{(D)})-\mathrm{Err}_{\mathrm{in}}{}'(h^{(D)})|>\tfrac{\varepsilon}{2}\}\mid B).$$

Das Ereignis $\{|\mathrm{Err}_{\mathrm{in}}(h^{(D)})-\mathrm{Err}_{\mathrm{in}}{}'(h^{(D)})|>\frac{\varepsilon}{2}\}$ umfasst wegen der besonderen Eigenschaft von $h^{(D)}$

das Ereignis $\{|\mathrm{Err_{in}}'(h^{(D)}) - \mathrm{Err_{out}}(h^{(D)})| \leq \frac{\varepsilon}{2}\}$ (man wende die Dreiecksungleichung an); daher ist

$$\mathbf{P}(\{|\mathrm{Err_{in}}(h^{(D)}) - \mathrm{Err_{in}}'(h^{(D)})| > \tfrac{\varepsilon}{2}\} \mid B) \geq \mathbf{P}(\{|\mathrm{Err_{in}}'(h^{(D)}) - \mathrm{Err_{out}}(h^{(D)})| \leq \tfrac{\varepsilon}{2} \mid B\}). \tag{3}$$

Hält man beim rechten Term in (3) die Stichprobe D fest und variiert nur D', ist man in der Situation der Hoeffding-Ungleichung für eine Hypothese (das Wahrscheinlichkeitsmaß ist auf den Raum der D' konzentriert). Sie liefert die Abschätzung

$$\mathbf{P}(\{|\mathrm{Err_{in}}'(h^{(D)}) - \mathrm{Err_{out}}(h^{(D)})| \leq \tfrac{\varepsilon}{2} \mid B\} \geq 1 - e^{-(1/2)\varepsilon^2 N}. \tag{4}$$

Insgesamt haben wir damit $\mathbf{P}(A \mid B) \geq (1 - 2e^{-(1/2)\varepsilon^2 N})$ gezeigt. □

Bemerkung 1 *Wir werden später das Resultat (2) unter der Bedingung* $1 - e^{-(1/2)\varepsilon^2 N} > \frac{1}{2}$ *anwenden. Dann lautet es einfach*

$$\mathbf{P}(B) \leq 2\,\mathbf{P}(A). \tag{5}$$

Satz 3 *([1], Lemma A.3.) Vorausgesetzt sei nun, dass eine* $2N$*-elementige Lernmenge aus* X *vorliegt. Mit* S *bezeichnen wir eine mögliche Aufteilung dieser Lernmenge in* D *und* D'*, sodass* $|D| = |D'| = N$ *ist. Alle Aufteilungen* S *seien gleichwahrscheinlich. Es gilt*

$$\mathbf{P}(\sup_{h \in H} |\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2}) \leq m_H(2N) \sup_S \sup_{h \in H} \mathbf{P}(|\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2}) \mid S). \tag{6}$$

Beweis

$$\begin{aligned}
\mathbf{P}(\sup_{h \in H} |\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2}) &= \sum_S \mathbf{P}(\sup_{h \in H} |\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2} \mid S)\,\mathbf{P}(S) \\
&\leq \sup_S \mathbf{P}(\sup_{h \in H} |\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2} \mid S) \\
&\leq \sup_S m_H(2N) \sup_{h \in H} \mathbf{P}(|\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2} \mid S).
\end{aligned}$$

Bei der letzten Ungleichung wurde darauf zurückgegriffen, dass es pro S nur $m_H(2N)$ verschiedene Werte für $|\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)|$ geben kann. □

Satz 4 *([1], Lemma A.4.) Mit den Bezeichnungen von Satz 3 seien* h *und* S *fix. Es gilt*

$$\mathbf{P}(|\mathrm{Err_{in}}(h) - \mathrm{Err_{in}}'(h)| > \tfrac{\varepsilon}{2} \mid S) \leq 2e^{-(1/8)\varepsilon^2 N}. \tag{7}$$

Beweis [1] greift auf folgendes Resultat (1963) von Hoeffding zurück:

Sei $A = \{a_1, \ldots, a_{2N}\} \subset [0, 1]$ und μ der Mittelwert von A. Sei $D = \{z_1, \ldots, z_N\} \subset A$ durch Ziehen aus A ohne Zurücklegen unter Gleichverteilungsannahme gewonnen worden. Sei ν der Mittelwert von D. Dann ist $\mathbf{P}(|\nu - \mu| > \varepsilon) \leq 2e^{-2\epsilon^2 N}$.

In unserem Fall ist $\mu = (\mathrm{Err}_{\mathrm{in}}(h) + \mathrm{Err}_{\mathrm{in}}{}'(h))/2$ und $\nu = \mathrm{Err}_{\mathrm{in}}(h)$. Zu beachten ist außerdem, dass $|\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{in}}{}'(h)| > \frac{\varepsilon}{2}$ äquivalent zu $|\mathrm{Err}_{\mathrm{in}}(h) - \mu| > \frac{\varepsilon}{4}$, denn μ liegt in der Mitte zwischen $\mathrm{Err}_{\mathrm{in}}(h)$ und $\mathrm{Err}_{\mathrm{in}}{}'(h)$. Es ist in unserem Fall also $\mathbf{P}(|\nu - \mu| > \frac{\varepsilon}{4}) \leq 2e^{-(1/8)\epsilon^2 N}$. □

2 Vapnik-Chervonenkis-Dimension

Die Wachstumsfunktion m_H gibt die maximale Anzahl von Dichotomien an, die eine bestimmte Hypothesenmenge H auf einer N-elementigen Teilmenge einer bestimmten Inputmenge X erzeugen kann; formal: $m_H(N) = \max_{\{x_1,\ldots,x_N\}\subset X} |H(x_1, \ldots, x_N)|$.

Definition 1 (VC-Dimension) *Seien Inputmenge 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 2 *Ist $d_{VC} < \infty$, so ist jedes $k > d_{VC}$ ein Breakpoint.*

Bemerkung 3 *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 4 *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 5 *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 $d_{VC} \leq d + 1$. □

Bemerkung 6 (Faustregel) *Für "vernünftige" ε's und rechte Seiten δ in (1) benötigt man Trainingsmengen vom Umfang $N \geq 10 d_{VC}$.*

Bemerkung 7 (Generalization bound) *Bezeichnet man die rechte Seite von (1) mit δ und löst man nach ε auf, bekommt man*

$$\varepsilon = \sqrt{\frac{8}{N} \ln \frac{4 m_H(2N)}{\delta}} =: \Omega(N, H, \delta).$$

Mit Wahrscheinlichkeit $1 - \delta$ ist

$$\mathrm{Err}_{\mathrm{out}} \leq \mathrm{Err}_{\mathrm{in}} + \Omega(N, H, \delta).$$

Zwar kann man mit einem komplexeren Hypothesenraum den in-sample error drücken, handelt sich aber dabei oft eine höhere VC-Dimension und ein größeres Ω ein.

Referenzen

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