Hoeffding-Ungleichung

M. Gruber

KW 50

Betrachten wir ein Klassifizierungsproblem mit Lernmenge $D = \{[x^{(1)}, y^{(1)}], \ldots, [x^{(N)}, y^{(N)}]\}$ und Hypothesenmenge H. Für die Hypothese $h \in H$ seien $U_1, \ldots, U_N$ unabhängige Kopien der Zufallsvariablen $\big[h(X) \neq Y\big]$. Der *in-sample error* $\mathrm{Err}_{\mathrm{in}}(h) = \frac{1}{N} \sum_{1 \leq i \leq N} \big[h(x^{(i)} \neq y^{(i)})\big]$ ist eine Realisierung der Zufallsvariablen $U = \frac{1}{N} \sum_{1 \leq i \leq N} U_i$, der *out-of-sample error* $\mathrm{Err}_{\mathrm{out}}(h)$ ist der Erwartungswert $\mathbf{E}\, U = \frac{1}{N} \sum_{1 \leq i \leq N} \mathbf{E}\, U_i = \mathbf{E}\, U_1 = \ldots = \mathbf{E}\, U_N$.

Die Größe $\sup_{h \in H} \big|\mathrm{Err}_{\mathrm{in}}(h) - \mathrm{Err}_{\mathrm{out}}(h)\big|$ ist der *Generalisierungsfehler* unseres Lernmodells. Er liegt zwischen 0 und 1. Im Folgenden geht es um die Frage, mit welcher Wahrscheinlichkeit der Generalisierungsfehler eine Toleranz $\varepsilon > 0$ überschreitet.

Betrachten wir zunächst ein einzelnes $V_i = U_i - \mathbf{E}\, U_i$. Es gilt $\mathbf{P}(V_i \in [a, b]) = 1$ mit $b - a = 1$ und $\mathbf{E}\, V_i = 0$. Für den Erwartungswert von e^{sV_i} haben gibt es eine interessante Abschätzung, wie man dem folgenden Satz entnehmen kann. Das Interessante an der Abschätzug ist, dass sie nicht von Verteilungseigenschaften von V_i abhängt (außer von Zentriertheit und Beschränktheit der Spannweite).

Satz 1 (Hoeffding's Lemma) *Für V mit $\mathbf{P}(V \in [a, b]) = 1$ und $\mathbf{E}\, V = 0$ gilt*

$$\mathbf{E}\, e^{sV} \leq e^{s^2 (b-a)^2 / 8}. \tag{1}$$

Beweis Es genügt, die Ungleichung für $b - a = 1$ zu zeigen. Der Ungleichung für den allgemeinen Fall bekommt man dann, indem man s durch $s(b - a)$ und V durch $V/(b - a)$ ersetzt.

Sei also $b - a = 1$. Wegen $\mathbf{E}\, V = 0 \in [a, b]$ ist somit $-1 \leq a \leq 0$. Betrachten wir die konvexe Funktion $v \mapsto e^{sv}$ auf dem Intervall $[a, b]$. Am linken Endpunkt nimmt sie den Wert e^{sa} an, am rechten e^{sb}. Dazwischen liegt sie unterhalb der Funktion, die e^{sa} mit e^{sb} linear verbindet:

$$e^{sv} \leq (b - v) e^{sa} + (v - a) e^{sb}.$$

Für die Zufallsvariable V bedeutet das

$$e^{sV} \leq (b - V) e^{sa} + (V - a) e^{sb}$$

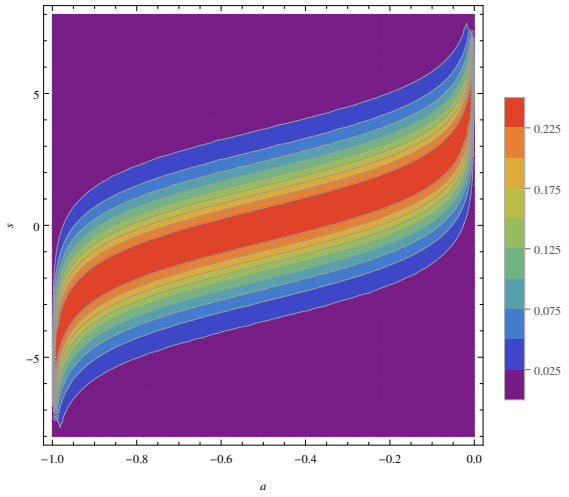
und für die Erwartungswerte

$$\mathbf{E}\, e^{sV} \leq (b - \mathbf{E}\, V) e^{sa} + (\mathbf{E}\, V - a) e^{sb}.$$

Da $\mathbf{E}\,V = 0$ und $b = a + 1$ ist, ist

$$\mathbf{E}\,e^{sV} \le be^{sa} - ae^{sb} = (a+1)e^{sa} - ae^{s(a+1)} = e^{sa}(a + 1 - ae^s).$$

Die rechte Seite $e^{sa}(a+1-ae^s)$ schreiben wir als $e^{L(s,a)}$ mit $L(s,a) = sa + \ln(a+1-ae^s)$. Wir führen eine Taylorentwicklung von $L(.,a)$ an der Stelle $s = 0$ durch: $L(s,a) = L(0,a) + D_1L(0,a)s + (1/2)D_1^2L(\tilde{s},a)s^2$. Die Differentiation liefert $D_1L(s,a) = a - ae^s/(a+1-ae^s)$ und $D_1^2L(s,a) = (a+1)(-ae^s)/(a+1-ae^s)^2$. Es ist $L(0,a) = 0$ und $D_1(0,a) = 0$. Einen Eindruck von $D_1^2L(s,a)$ vermittelt der Contourplot für $-1 \le a \le 0$ und $-8 \le s \le 8$:



Der Ausdruck $(a+1)(-ae^s)/(a+1-ae^s)^2$ ist von der Form $xy/(x+y)^2$ mit $x \ge 0$ und $y \ge 0$ bzw. (in Polarkoordinaten) $h(t) = r^2 \cos\pi t \sin\pi t/(r\cos\pi t + r\sin\pi t)^2 = \cos\pi t \sin\pi t/(1 + 2\cos\pi t \sin\pi t)$ mit $0 \le t \le 1/2$.

Die Funktion h nimmt bei $t = 1/4$ ihr Maximum an. Der Wert des Maximums ist $1/4$. Damit wissen wir, dass $|D_1^2L(s,a)| \le 1/4$ ist und können den Term zweiter Ordnung in der Taylorentwicklung nach oben abschätzen: $L(s,a) \le s^2/8$. □

Nun betrachten wir $\bar{V} = \frac{1}{N}\sum_{1\le i\le N} V_i$.

Satz 2 (Hoeffding-Ungleichung) *Seien $V_i, i = 1,\dots,N$ unabhängige Zufallsvariablen mit $\mathbf{P}(V_i \in [0,1]) = 1$ und sei $\bar{V} = \frac{1}{N}\sum_{1\le i\le N} V_i$. Dann gilt*

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

Beweis Für jedes $s \in \mathbf{R}$ ist $\mathbf{P}(\bar{V} - \mathbf{E}\,\bar{V} \ge \varepsilon) = \mathbf{P}(e^{s(\bar{V}-\mathbf{E}\,\bar{V})} \ge e^{s\varepsilon})$. Wir werden s noch geeignet bestimmen.

Es ist $\mathbf{E}\,e^{s(\bar{V}-\mathbf{E}\,\bar{V})} \ge e^{s\varepsilon}\,\mathbf{P}(e^{s(\bar{V}-\mathbf{E}\,\bar{V})} \ge e^{s\varepsilon})$, folglich $\mathbf{P}(e^{s(\bar{V}-\mathbf{E}\,\bar{V})} \ge e^{s\varepsilon}) \le e^{-s\varepsilon}\,\mathbf{E}\,e^{s(\bar{V}-\mathbf{E}\,\bar{V})}$.

Wegen der Unabhängigkeit der Summanden von $\bar{V}$ ist $\mathbf{E}\,e^{s(\bar{V}-\mathbf{E}\,\bar{V})} = \mathbf{E}\,e^{(s/N)\sum_i (V_i - \mathbf{E}\,V_i))} = \prod_i \mathbf{E}\,e^{(s/N)(V_i-\mathbf{E}\,V_i)}$.

Jedes $V_i - \mathbf{E}\,V_i$ erfüllt die Voraussetzung von Hoeffding's Lemma mit $b - a = 1$. Jeden Faktor $\mathbf{E}\,e^{(s/N)(V_i-\mathbf{E}\,V_i)}$ kann man daher nach oben abschätzen durch $e^{s^2/(8N^2)}$.

Insgesamt ist damit $e^{-s\varepsilon}\,\mathbf{E}\,e^{s(\bar{V}-\mathbf{E}\,\bar{V})} \le e^{-s\varepsilon}e^{s^2/(8N)} = e^{-s\varepsilon + s^2/(8N)}$.

Wir bestimmen nun s so, dass diese obere Schranke minimal wird. Das Minimum des Exponenten wird angenommen an der Stelle s, an der $-\varepsilon + s/(4N) = 0$ ist. Das ist die Stelle $s = 4\varepsilon N$. Setzt man dieses s in den Exponenten von $e^{-s\varepsilon+s^2/(8N)}$ ein, erhält man den Wert $e^{-2\epsilon^2 N}$. □

Bemerkung 1 *Aus Gleichung (2) folgt für eine Hypothese $h \in H$*

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

Bemerkung 2 *Aus Gleichung (3) folgt für einen endlichen Hypothesenraum $H = \{h_1, \ldots, h_M\}$*

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