

5 Wieviele Beispiele brauchen wir um unendlich viele Hypothesen zu unterscheiden?

5.1 Die Vapnik-Chervonenkis-Dimension (VC)

Wir kommen zurück auf's PAC-Modell. Zur Erinnerung:

PAC = Probably Approximately Correct

Ziel: Lerne aus Beispielen, so dass mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ der Fehler beim Klassifizieren höchstens ϵ ist.

Ist D eine W'keitsverteilung auf Ω und $C \in \mathcal{C}$ das zu lernende Konzept, so können wir den Fehler der Hypothese $H \in \mathcal{H}$ durch

$F_D(C, H) := D((C \cup H) \setminus (C \cap H))$ definieren.

Fairness-Bedingung: Bewerte den Fehler mit derselben Verteilung D , gemäß der die Lernbeispiele erzeugt wurden.

PAC-Algorithmus für eine Konzeptklasse \mathcal{C}

Ablauf:

- Eingabe: Vertrauensparameter δ , Fehlerparameter ϵ , Länge n der Beispiele
- Bestimme Anzahl $s = s(\delta, \epsilon, n)$ der anzufordernden Beispiele.
- Lies s Beispiele an
- Gib ein $H \in \mathcal{H}$ aus.

Kriterium: Für alle $\epsilon, \delta \in (0, 1]$, $C \in \mathcal{C}$ und alle Verteilungen D auf Ω gilt:

$$W_{sD}(F_D(C, H) \leq \epsilon) \geq 1 - \delta$$

($W_{sD}(\cdot)$:= W'keit unter der Voraussetzung, dass die Lerndaten aus D kommen)

Ein PAC-Algorithmus für endliches $|\mathcal{H}|$:

Falls $\mathcal{H} = \mathcal{C}$ und es leicht ist, für jede mit $C \in \mathcal{C}$ verträgliche Folge von Beispielen $\{(x_1, b_1), \dots, (x_s, b_s)\} \subset \Omega \times \{0, 1\}$ (d.h. $b_i = 1 \Leftrightarrow x_i \in C$) eine konsistente Hypothese $H \in \mathcal{H}$ zu finden (d.h. $b_i = 1 \Leftrightarrow x_i \in H$), dann

- setze $s := \lceil (\ln(|\mathcal{C}|) - \ln(\delta)) / \varepsilon \rceil$
- lies s Beispiele ein
- Gib dazu konsistente Hypothese H aus

Brauchen wir also unendlich viele Beispiele falls $|\mathcal{H}| = \infty$?

Definition: Die Konzeptklasse \mathcal{C} **zertrümmert** die Menge S , falls

$$\mathcal{P}(S) = \{S \cap C \mid C \in \mathcal{C}\}.$$

Die **Vapnik-Chervonenkis-Dimension** $VC(\mathcal{C})$ von \mathcal{C} ist die Mächtigkeit der größten Menge S , die von \mathcal{C} zertrümmert wird.

Beispiele:

- Die Klasse der Halbräume im \mathbb{R}^n hat VC-Dim $n + 1$
- Die Klasse der achsenparallelen Rechtecke in $\mathbb{Z}^d \cap [-2^n, 2^n]$ hat VC-Dim $2d$.
- Die Klasse der $U \subset \mathbb{R}$ mit $\exists \alpha : U = \{x : \sin(x\alpha) > 0\}$ hat VC-Dim ∞ , denn sie zerschmettert die Menge $\{10^{-i} \mid i = 1, \dots, n\}$ für jedes n .

Satz 15 Sei $2 \leq VC(\mathcal{C}) < \infty$, $\varepsilon < 1/4$ und $\delta < 1/2$. Dann muss der entsprechende PAC-Algorithmus mindestens

$$\Omega\left(\frac{1}{\varepsilon} \left[VC(\mathcal{C}) + \ln\left(\frac{1}{\delta}\right) \right]\right)$$

Beispiele anfordern.

Beweis: Zeige zunächst $\Omega(\text{VC}(\mathcal{C})/\varepsilon)$.

Konstruiere schwierige Verteilung \mathcal{D} :

\mathcal{C} mit $\text{VC}(\mathcal{C}) = d + 1$ zertrümmere $S = (x_0, \dots, x_d)$.

$$\mathcal{D}(x) = \begin{cases} 0 & \text{für } x \notin S \\ 1 - 4\varepsilon & \text{für } x = x_0 \\ 4\varepsilon/d & \text{für } x \in S \setminus x_0 \end{cases}$$

\Rightarrow Konzeptklasse $\{S \cap C \mid C \in \mathcal{C}\}$.

Bei weniger als $d/(16\varepsilon)$ Bsp. sehen wir (wahrscheinlich) mehr als $d - 4\varepsilon d/(16\varepsilon) = 3d/4$ Bsp. nicht und klassifizieren davon (wahrscheinlich) mehr als $3d/8$ falsch.

\Rightarrow Fehler $\geq 3d/8 \cdot 4\varepsilon/d = 3/2\varepsilon$.

Wir zeigen noch, dass mehr als $\Omega(\ln(1/\delta)/\varepsilon)$ Beispiele nötig sind:

Wir betrachten ein D welches auf ein $x \in X$ W 'keit 2ε legt. Die W 'keit, x nicht in den Beispielen zu sehen, sollte $< 3\delta$ sein, sonst wird es mit W 'keit $> \delta$ falsch klassifiziert. Bei n Versuchen, sehen wir es *nicht* mit W 'keit $(1 - 2\varepsilon)^n$. Also soll gelten $(1 - 2\varepsilon)^n \leq 3\delta$ Logarithmieren und Auflösen ergibt:

$$n \geq \frac{\ln(3\delta)}{\ln(1 - 2\varepsilon)} = \frac{-\ln(3\delta)}{-\ln(1 - 2\varepsilon)} = \frac{\ln(\frac{1}{3\delta})}{\ln\left(\frac{1}{1-2\varepsilon}\right)} = \frac{\ln(\frac{1}{3\delta})}{\ln\left(1 + \frac{\varepsilon}{\frac{1}{2}-\varepsilon}\right)}$$

Da \ln konvex ist und im Punkt 1 die Steigung 1 hat, gilt $\ln\left(1 + \frac{\varepsilon}{\frac{1}{2}-\varepsilon}\right) < \frac{\varepsilon}{\frac{1}{2}-\varepsilon}$.

Also folgt:

$$n \geq \frac{\ln(\frac{1}{3\delta})}{\frac{\varepsilon}{\frac{1}{2}-\varepsilon}} > 4 \frac{\ln(\frac{1}{\delta}) - \ln 3}{\varepsilon} = \Omega(\ln(1/\delta)/\varepsilon)$$

□

Satz 16 *Es gibt einen PAC-Algorithmus, der höchstens*

$$\left\lceil \frac{4}{\varepsilon} \cdot \left(VC(\mathcal{C}) \cdot \ln \left(\frac{12}{\varepsilon} \right) + \ln \left(\frac{2}{\delta} \right) \right) \right\rceil$$

anfordert.

Beweisschritte: S sei Beispielmenge gemäß Verteilung D .

Menge der ε -Fehlerregionen von $C \in \mathcal{C}$ sei

$$\Delta_\varepsilon(C) := \{C \cup C' \setminus C \cap C' \mid C' \in \mathcal{C}, F_{\mathcal{D}}(C, C') > \varepsilon\}$$

$$A := \{\exists U \in \Delta_\varepsilon(C) : S \cap U = \emptyset\}$$

zu zeigen: $\Pr(A) \leq \delta$

mit Hilfe von Gedankenexperiment: Sei S' eine weitere Beispielfolge mit $|S'| = |S| =: s$.

$$B := \{\exists U \in \Delta_\varepsilon(C) : S \cap U = \emptyset \wedge |U \cap S'| \geq \varepsilon \cdot s/2\}$$

Zeige mit Hilfe von Chernoff-Schranke:

Für $s \geq 8/\varepsilon$ gilt $\Pr(B) = \Pr(A) \cdot \Pr(B|A) \geq \Pr(A) \cdot \left(1 - e^{-\left(\frac{1}{2}\right)^2 \frac{\varepsilon \cdot s}{2}}\right) \geq \frac{1}{2} \Pr(A)$

Dann bleibt zu zeigen: $\Pr(B) \leq \delta/2$.

$$\begin{aligned}
\Pr(B) &\leq \sum_U \Pr\left(U \cap S = \emptyset \wedge |U \cap S'| \geq \frac{\varepsilon \cdot s}{2}\right) \\
&\leq \mathbb{E}|\{U = (S \cup S') \cap C \mid C \in \mathcal{C}\}| \cdot \left(\frac{1}{2}\right)^{\frac{\varepsilon s}{2}}
\end{aligned}$$

Um daraus $\Pr(B) \leq \delta/2$ zu zeigen, verwenden wir folgendes

Lemma 1

$$|\{C \cap S \mid C \in \mathcal{C}\}| \leq \sum_{i=0}^{\text{VC}(\mathcal{C})} \binom{|S|}{i} \begin{cases} = 2^{|S|} & \text{für } |S| < \text{VC}(\mathcal{C}) \\ \leq \left(\frac{e \cdot |S|}{\text{VC}(\mathcal{C})}\right)^{\text{VC}(\mathcal{C})} & \text{für } |S| \geq \text{VC}(\mathcal{C}) \geq 1 \end{cases}$$

Damit folgt nämlich $\Pr(A) \leq 2 \Pr(B) \leq 2 \cdot \left(\frac{e \cdot 2s}{\text{VC}(\mathcal{C})}\right)^{\text{VC}(\mathcal{C})} \cdot e^{-\varepsilon s/2}$.

Bleibt also zu zeigen, dass gilt $\left(\frac{e \cdot 2s}{\text{VC}(\mathcal{C})}\right)^{\text{VC}(\mathcal{C})} \cdot e^{-\varepsilon s/2} \leq \delta/2$

oder äquivalent:

$$\text{VC}(\mathcal{C}) \cdot \ln \left(\frac{e \cdot 2s}{\text{VC}(\mathcal{C})} \right) - \frac{\varepsilon s}{2} \leq \ln \frac{\delta}{2}$$

Nach Anwendung diverser Umformungen und Ungleichungen für Logarithmen zeigt sich, dass dafür folgendes genügt:

$$s \geq \frac{4}{\varepsilon} \cdot \left(\text{VC}(\mathcal{C}) \cdot \ln \left(\frac{12}{\varepsilon} \right) + \ln \left(\frac{2}{\delta} \right) \right)$$

□

Beweis des Lemmas: Sei $\Pi_{\mathcal{C}}(S) := \{C \cap S \mid C \in \mathcal{C}\}$.

erste Ungleichung: Induktion über $|S|$: Sei $|S| = m$ und die Induktionsannahme ist, dass die Aussage

$$|\Pi_{\mathcal{C}}(S')| \leq \sum_{i=0}^{\text{VC}(\mathcal{C})} \binom{|S'|}{i}$$

für alle S' mit $|S'| < m$ gilt, insbesondere also für $S' := S \setminus \{x\}$ (für irgendein gewähltes $x \in S$).

Sei

$$\mathcal{C}' := \{U \in \Pi_{\mathcal{C}}(S) \mid x \notin U, U \cup \{x\} \in \Pi_{\mathcal{C}}(S)\}$$

Dann gilt $\text{VC}(\mathcal{C}') < \text{VC}(\mathcal{C})$ (wegen $\text{VC}(\Pi_{\mathcal{C}}(S)) < \text{VC}(\mathcal{C})$) und

$$\begin{aligned} |\Pi_{\mathcal{C}}(S)| &= |\Pi_{\mathcal{C}}(S \setminus \{x\})| + |\Pi_{\mathcal{C}'}(S \setminus \{x\})| \\ &\leq \sum_{i=0}^{\text{VC}(\mathcal{C})} \binom{|S| - 1}{i} + \sum_{i=0}^{\text{VC}(\mathcal{C}) - 1} \binom{|S| - 1}{i} \\ &= \sum_{i=0}^{\text{VC}(\mathcal{C})} \binom{|S|}{i} \end{aligned}$$

zweite Ungleichung: Für $m < d$ gilt $\sum_{i=0}^d \binom{m}{i} = \sum_{i=0}^m \binom{m}{i} = 2^m$, für $m \geq d \geq 1$ hingegen:

$$\begin{aligned} \left(\frac{d}{m}\right)^d \cdot \sum_{i=0}^d \binom{m}{i} &\leq \sum_{i=0}^m \left(\frac{d}{m}\right)^i \cdot \binom{m}{i} \\ &= \left(1 + \frac{d}{m}\right)^m \\ &\leq e^d \end{aligned}$$

und damit

$$\sum_{i=0}^d \binom{m}{i} \leq \left(\frac{e \cdot m}{d}\right)^d$$

□

5.2 Hard Margins

Als Analogon zur VC-Dimension definieren wir:

Definition Sei \mathcal{F} eine Familie von Funktionen $f : X \rightarrow \mathbb{R}$ und $\gamma > 0$. Die Menge $S = \{x_1, \dots, x_m\} \subset X$ wird **γ -zertrümmert (*shattered*) durch \mathcal{F}** , falls es einen Vektor r gibt, so dass für jeden “Klassifizierungsvektor” $b \in \{-1, 1\}^m$ eine Funktion $f_b \in \mathcal{F}$ existiert, so dass gilt

$$f_b(x_i) \begin{cases} \geq r_i + \gamma & \text{falls } b_i = 1 \\ \leq r_i - \gamma & \text{falls } b_i = -1 \end{cases}$$

Die **Fat-Shattering-Dimension $\text{fat}_{\mathcal{F}}(\gamma)$ von \mathcal{F} für γ** ist dann das Maximum von $|S|$ über alle S , die durch \mathcal{F} γ -zertrümmert werden.

Definition Sei \mathcal{F} eine Familie von Funktionen $f : X \rightarrow \mathbb{R}$ und $\gamma > 0$. Eine **γ -Überdeckung** von \mathcal{F} für $S = \{x_1, \dots, x_m\} \subset X$ ist eine endliche Familie $B(S)$ von Funktionen mit:

$$\forall f \in \mathcal{F} \exists g_f \in B(S) \forall x \in S : |f(x) - g_f(x)| < \gamma$$

Die **Überdeckungszahl** $\mathcal{N}(\mathcal{F}, S, \gamma)$ von \mathcal{F} für S ist die Größe der kleinstmöglichen γ -Überdeckung und die **Überdeckungszahl** $\mathcal{N}(\mathcal{F}, m, \gamma)$ von \mathcal{F} ist das Maximum von $\mathcal{N}(\mathcal{F}, S, \gamma)$ über alle m -elementigen $S \subset X$.

Satz 17 Sei \mathcal{F} eine Menge von Funktionen $f : X \rightarrow \mathbb{R}$ und $\gamma > 0$. Für jede Wahrscheinlichkeitsverteilung D auf $X \times \{-1, 1\}$ gilt für jede Hypothese $f \in \mathcal{F}$ mit Margin $m_S(f) \geq \gamma$ auf einer gemäß D gezogenen Beispiele-Menge S mit $|S| > 2/\varepsilon$ mit Wahrscheinlichkeit $1 - \delta$, dass der Klassifikationsfehler $F_D(f)$ kleiner gleich

$$\frac{2}{|S|} \left(\log \mathcal{N}(\mathcal{F}, 2|S|, \gamma/2) + \log \frac{2}{\delta} \right)$$

ist.

Beweis Analog zum Beweis von Satz 16 mit

$$A := \{\exists f \in \mathcal{F} : F_D(\text{sign}(f)) > \varepsilon, m_S(f) \geq \gamma\}$$

und

$$B := A \wedge \{\text{sign}(f) \text{ missklassifiziert mehr als } \varepsilon m/2 \text{ von } S'\}.$$

□

Analog zu Lemma 1 gilt folgendes Lemma, mit dem wir die Brücke zwischen der Überdeckungsanzahl und der fat-shattering-Dimension schlagen:

Lemma 2 *Sei \mathcal{F} eine Familie von Funktionen $X \rightarrow [a, b]$ und D eine Verteilung auf X . Dann gilt für $0 < \gamma < 1$:*

$$\log \mathcal{N}(\mathcal{F}, l, \gamma) \leq 1 + \text{fat}_{\mathcal{F}}(\gamma/4) \cdot \log \frac{2el(b-a)}{\text{fat}_{\mathcal{F}}(\gamma/4) \cdot \gamma} \log \frac{4l(b-a)^2}{\gamma^2}$$

Daraus erhalten wir:

Satz 18 \mathcal{F} bestehe aus Funktionen $X \rightarrow [0, 1]$ und es seien $\gamma > 0$ und eine Verteilung D auf $X \times \{-1, 1\}$ gegeben, nach der eine Stichprobe S mit mehr als $2/\varepsilon$ Elementen zusammengestellt sei. Sei f eine Hypothese mit Margin $m_S(f) \geq \gamma$ und es sei $\text{fat}_{\mathcal{F}}(\gamma/8) < |S|$. Dann ist mit Wahrscheinlichkeit $1 - \delta$ der Fehler von f kleiner gleich:

$$\frac{2}{|S|} \cdot \left(\text{fat}_{\mathcal{F}}(\gamma/8) \cdot \log \frac{8e|S|}{\text{fat}_{\mathcal{F}}(\gamma/8) \cdot \gamma} \cdot \log \frac{32|S|}{\gamma^2} + \log \frac{4}{\delta} \right)$$

Um das auf SVM anzuwenden, benötigen wir die fat-shattering-Dimension linearer Funktionen-Klassen – oder zumindest eine Abschätzung:

Satz 19 Sei $R > 0$ und $X := \{x \in V : \|x\| \leq R\}$ und sei

$$\mathcal{L} = \{X \rightarrow \mathbb{R}, x \mapsto \langle w, x \rangle : \|w\| \leq 1\}$$

Dann gilt

$$\text{fat}_{\mathcal{L}}(\gamma) \leq \left(\frac{R}{\gamma}\right)^2$$

Beweis Wird S durch \mathcal{L} γ -zertrümmert, so gilt für alle $T \subset S$:

$$\left\| \sum_{x \in T} x - \sum_{y \in S \setminus T} y \right\| \geq |S| \cdot \gamma.$$

Gilt $\|x\| \leq R$ für alle $x \in S$, so folgt bei rein zufälliger Wahl von T :

$$\mathbb{E} \left[\left\| \sum_{x \in T} x - \sum_{y \in S \setminus T} y \right\|^2 \right] \leq R^2 \cdot |S|$$

Also gibt es mindestens ein T mit $\left\| \sum_{x \in T} x - \sum_{y \in S \setminus T} y \right\| \leq R \cdot \sqrt{|S|}$ und mit $|S| = \text{fat}_{\mathcal{L}}(\gamma)$ folgt

$$\gamma \cdot \text{fat}_{\mathcal{L}}(\gamma) \leq \left\| \sum_{x \in T} x - \sum_{y \in S \setminus T} y \right\| \leq R \cdot \sqrt{\text{fat}_{\mathcal{L}}(\gamma)}$$

□

5.3 Soft Margins

Slack von $(x, b) \in X \times \{-1, 1\}$ bzgl. $f : X \rightarrow \mathbb{R}$ und Margin γ :

$$s(x, b, \gamma, f) := \max\{0, \gamma - b \cdot f(x)\}$$

für $S := \{(x_1, b_1), (x_2, b_2), \dots, (x_m, b_m)\}$:

$$s(S, \gamma, f) := (s(x_1, b_1, \gamma, f), s(x_2, b_2, \gamma, f), \dots, s(x_m, b_m, \gamma, f))$$

Satz 20 Für die aus m Beispielen gelernte lineare threshold-Klassifikation $f : x \mapsto \langle x, \beta \rangle + \beta_0$ mit $\|\beta\| = 1$ und $\gamma > 0$ und jede W 'keits-Verteilung D auf $\{x \in X \mid \|x\| \leq R\} \times \{-1, 1\}$ ist der Fehler mit W 'keit $1 - \delta$ in

$$O \left(\frac{1}{m} \left(\frac{R^2 + \|s(S, \gamma, f)\|^2}{\gamma^2} \log^2 m + \log \frac{1}{\delta} \right) \right)$$

Beweisidee:

$$L(X) := \{g : X \rightarrow \mathbb{R} \mid g(x) \neq 0 \text{ nur für abzählbar viele } x, \sum_x g(x)^2 < \infty\}$$

$$\langle g, h \rangle := \sum_x g(x) \cdot h(x)$$

$$H : X \times L(X) \rightarrow \mathbb{R}, (x, f) \mapsto f(x) + \left\langle \sum_j s(x_j, b_j, \gamma, f) \cdot b_j \cdot \delta_{x_j}, \delta_x \right\rangle$$

mit $\delta_x(y) = 1$ für $x = y$ und 0 sonst.

Verwende als Beispiele $((x_i, \delta_{x_i}), b_i)$.

Damit gilt $b_i \cdot H(x_i, \delta_{x_i}) \geq \gamma$ und $H(x, h) = h(x)$ für nicht-Beispiele x .

wende hard-Margin-Sätze auf H an.