Herleitung oberer Schranken für die Approximation mit radialen Basisfunktionen unter Verwendung von Beweistechniken der statistischen Lerntheorie

Kolloquium zur Masterarbeit an der Universität Bayreuth

Dominik Köhler

Bayreuth, October 10, 2023

Einführung

Notation

$Z \subset \mathbb{R}^k$, $\mathcal{F} \subset \{f : Z \to [A, B] \text{ messbar}\}$, $\mathcal{M} \subset \{\phi : \mathbb{R}_0^+ \to [A, B] \text{ messbar}\}$,
$z_1, \ldots, z_n \in Z$, $\lambda \in \mathcal{L}_1(Z) \geq 0, \|\lambda\|_1 = 1$, $\delta \in (0, 1)$

Ziel: Finde eine obere Schranke für die gleichmäßige Konvergenz von:

$$\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \right| \leq \alpha_{n,\mathcal{F}}, \quad \alpha_{n,\mathcal{F}} \overset{n\to\infty}{\longrightarrow} 0$$

Methode: Nutze Schranken aus der statistischen Lerntheorie:

$$P\left\{ z_1, \ldots, z_n \in Z : \sup_{f \in \mathcal{F}} \left| E[f(z)] - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \right| \leq \alpha_{n,\delta,\mathcal{F}} \right\} \geq 1 - \delta, \quad \alpha_{n,\delta,\mathcal{F}} \overset{n\to\infty}{\longrightarrow} 0$$

Anwendung auf radiale Kerne:

$$\sup_{\phi \in \mathcal{M}, t \in Z} \left| \int_z \phi(\|z - t\|_2)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} \phi(\|z_i - t\|_2) \right| \leq C(B - A)\sqrt{\frac{c_{\mathcal{M}}(2k + 3)}{n}}$$

Gliederung

1. Statistische Lerntheorie
   - Die VC-Dimension
2. Die Abschätzung
3. Die VC-Dimension von RBF
4. Resultat

Definitionen

Notation

$$X \subset \mathbb{R}^k, \qquad Y \subset \mathbb{R}, \qquad P \text{ unbekanntes W'Maß auf } X \times Y,$$
$$(x_1, y_1), \ldots, (x_n, y_n), (x, y) \in (X \times Y) \text{ i.i.d. gemäß } P \text{ verteilt}$$

- Ziel: Finde Funktion f, welche die verteilten Punkte möglichst gut beschreibt, also:
$$(x, f(x)) \approx (x, y).$$
- Für die Bestimmung des Fehlers nutzen wir eine *Verlustfunktion*:
$$L(x, y, f(x)) : X \times Y \times \mathbb{R} \to \mathbb{R}_0^+, \qquad L \text{ messbar},$$
und erhalten das *empirische Risiko* auf den Daten $(x_1, y_1), \ldots, (x_n, y_n)$:
$$\widehat{R}(f) := \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f(x_i))$$
und für den Fehler auf ganz $X \times Y$ das *Risiko*:
$$R(f) := \int_{X \times Y} L(x, y, f(x)) dP(x, y)$$

Frage

Lässt sich das empirische Risiko auf den Daten $(x_1, y_1), \ldots, (x_n, y_n)$ auf das Risiko über ganz $X \times Y$ verallgemeinern?

Antwort

Dafür bestimmen wir den Abstand zwischen $R(f)$ und $\widehat{R}(f)$:

$$P\left\{(x_1, y_1), \ldots, (x_n, y_n) \in X \times Y : \left|R(f) - \widehat{R}(f)\right| \leq \alpha_{n,\delta,f}\right\} \geq 1 - \delta, \quad \alpha_{n,\delta,f} \overset{n\to\infty}{\longrightarrow} 0$$

Für eine feste Funktion f gilt dies nach dem Gesetz der großen Zahlen.

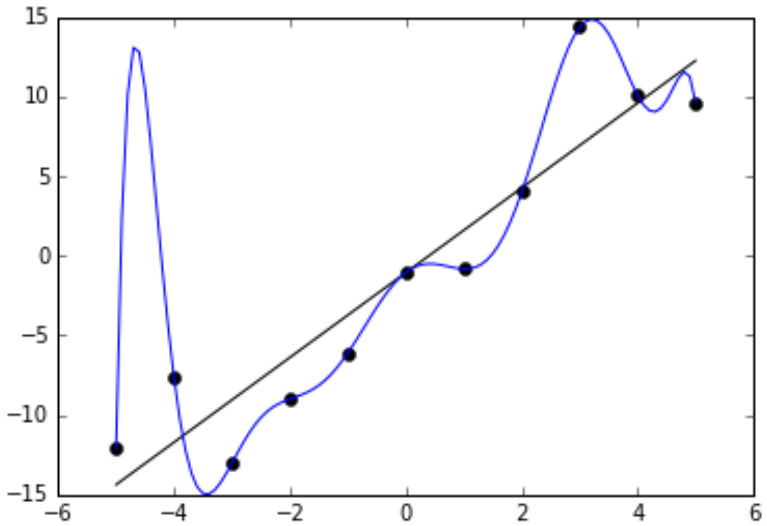
Abschätzung des Risikos

Frage

Lässt sich das empirische Risiko auf den n beobachteten Daten für alle Funktionen $f \in \mathcal{F} \subset \{f : X \to Y \text{ messbar}\}$ auf das Risiko verallgemeinern, also:

$$P\left\{(x_1, y_1), \ldots, (x_n, y_n) \in X \times Y : \sup_{f \in \mathcal{F}} \left|R(f) - \widehat{R}(f)\right| \leq \alpha_{n,\delta,\mathcal{F}}\right\} \geq 1 - \delta?$$

Im allgemeinen nicht:



Um die Menge $\mathcal{F}$ einzuschränken, nutzen wir *Komplexitätsmaße*.

Vereinfachte Voraussetzungen

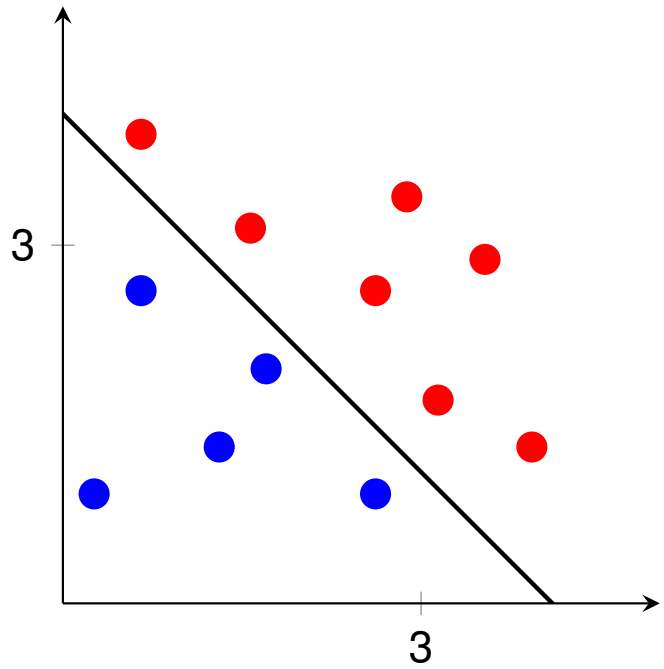
Wir ersetzen:

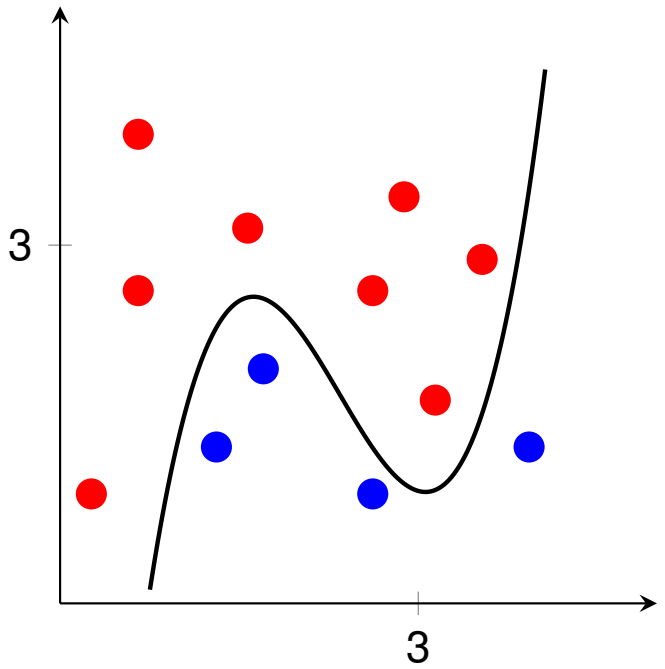
$$
\begin{aligned}
X \times Y &\mapsto Z \subset \mathbb{R}^k \\
L(x, y, f(x)) &\mapsto f(z), \qquad f : Z \to \mathbb{R} \\
P &\mapsto \lambda(z), \qquad \lambda \in \mathcal{L}_1(Z) \geq 0, \|\lambda\|_1 = 1 \\
R(f) = \int_{X \times Y} L(x, y, f) dP(x, y) &\mapsto \int_Z f(z)\lambda(z)dz \\
\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} L(x_i, y_i, f) &\mapsto \frac{1}{n} \sum_{i=1}^{n} f(z_i)
\end{aligned}
$$

Beispiel zur Klassifikation

Notation

$S = \{z_1, \ldots, z_n\} \subset Z \subset \mathbb{R}^k, \qquad \mathcal{F}_{0,1} \subset \{f : Z \to \{0,1\} \text{ messbar}\}$





Reellwertige Funktionen für die Klassifikation

Notation

$$Z \subset \mathbb{R}^k, \quad \mathcal{F} \subset \{f : Z \to \mathbb{R}\}, \quad \mathcal{H} : \mathbb{R} \to \{0,1\}, f(z) \mapsto \begin{cases} 1, & f(z) \geq 0, \\ 0, & f(z) < 0. \end{cases}$$

Definition

- *Induzieren: $S^+ \subset S$ wird durch $\mathcal{F}$ induziert, falls $f \in \mathcal{F}$ und $\beta \in \mathbb{R}$ existiert, sodass:*

$$\begin{cases} z \in S^+ & \implies \mathcal{H}(f(z) - \beta) = 1 \\ z \in S \backslash S^+ & \implies \mathcal{H}(f(z) - \beta) = 0 \end{cases}$$

- *Splittern: S wird durch $\mathcal{F}$ gesplittert, falls alle 2^n Teilmengen von S durch $\mathcal{F}$ induziert werden.*

Frage

Wie viele Teilmengen $S^+ \subset S$ lassen sich durch ein $f \in \mathcal{F}$ und $\beta \in \mathbb{R}$ induzieren, also

$$\exists f \in \mathcal{F}, \beta \in \mathbb{R} : \forall z \in S : \quad z \in S^+ \Leftrightarrow \mathcal{H}(f(z) - \beta) = 1 \ ?$$

Diese zählen wir:

$$\mathcal{G}_{\mathcal{F},S} := \text{ die Anzahl der von } \mathcal{F} \text{ induzierten Teilmengen von } S$$

Die Vapnik-Chervonenkis-Dimension

Notation

$\mathcal{G}_{\mathcal{F},S} :=$ *die Anzahl der von* $\mathcal{F}$ *induzierten Teilmengen von S,* $\mathcal{G}_{\mathcal{F},S} \leq 2^n$

Definition

Die Wachstumsfunktion von $\mathcal{F}$ *definieren wir als:*

$$\mathcal{G}_{\mathcal{F}}(n) := \max_{S \subset Z, \#S = n} \mathcal{G}_{\mathcal{F},S}(n)$$

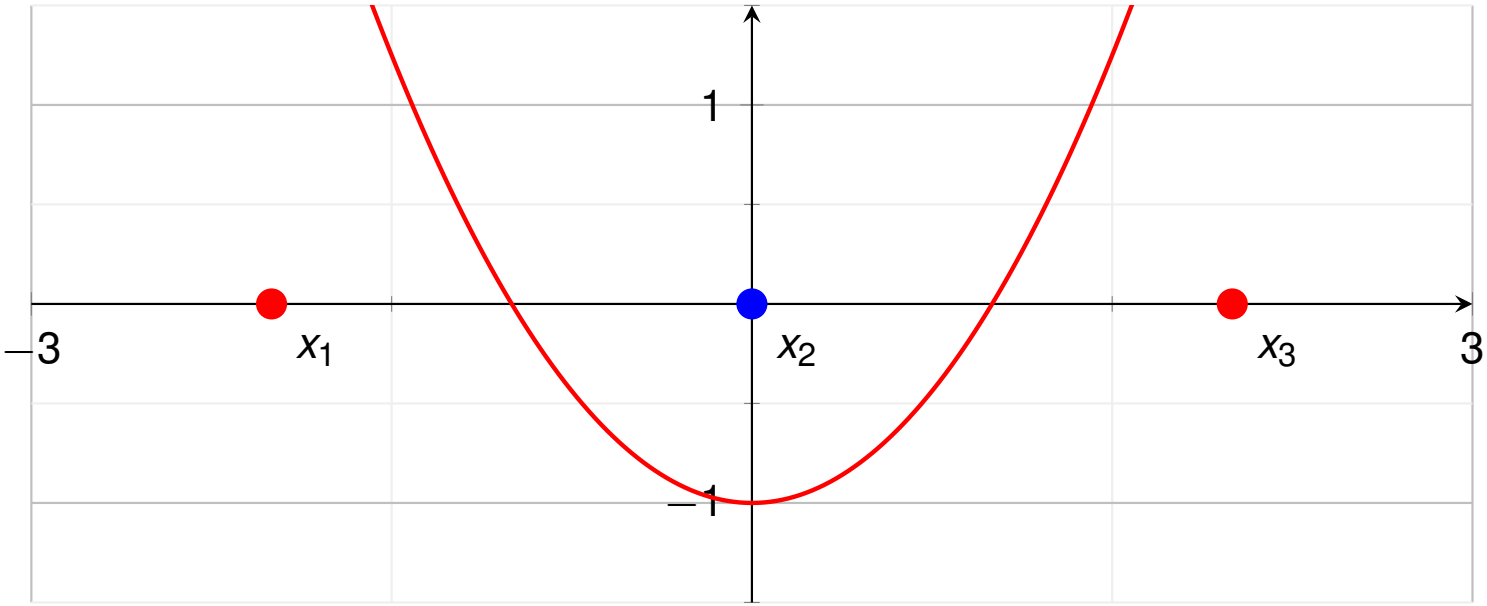
Definition

$$\mathrm{VCD}(\mathcal{F}) := \begin{cases} \max\{n \in \mathbb{N}_0 \mid \mathcal{G}_{\mathcal{F}}(n) = 2^n\}, & \textit{falls existent,} \\ \infty, & \textit{sonst.} \end{cases}$$

$\mathrm{VCD}(\mathcal{F}) \geq n \Leftrightarrow \exists S \subset Z, \#S = n$: alle 2^n Teilmengen von S werden induziert.

Beispiel: nach oben geöffnete Parabeln

$$\mathcal{F} := \{f : \mathbb{R} \to \mathbb{R}, x \mapsto (x-a)(x-b), \quad a, b \in \mathbb{R}\}$$



- $\{x_2\} \subset \{x_1, x_2, x_3\}$ nicht durch $\mathcal{F}$ induziert $\implies \text{VCD}(\mathcal{F}) < 3$
- $\{x_1, x_3\}$ durch $\mathcal{F}$ gesplittert $\implies \text{VCD}(\mathcal{F}) \geq 2$

Insgesamt:

$$\text{VCD}(\mathcal{F} = 2)$$

Abschätzung

Notation

$\mathcal{F} \subset \{f : Z \to [A, B]\}$, $d = \mathrm{VCD}(\mathcal{F})$, *C Konstante unabhängig von* $\mathcal{F}$,
$\lambda \in \mathcal{L}_1(Z) \geq 0, \|\lambda\|_1 = 1$.

Es gilt:

- Für alle $\delta \in (0, 1)$ gilt mit einer Wahrscheinlichkeit von mindestens $1 - \delta$:

$$\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i) \right| \leq C(B - A)\frac{1}{\sqrt{n}} \max\left\{ \sqrt{d}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}.$$

- Es existieren Punkte $\tilde{z}_1, \ldots, \tilde{z}_n \in Z$

$$\sup_{f \in \mathcal{F}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i) \right| \leq C(B - A)\sqrt{\frac{d}{n}}.$$

VC-Dimension für radiale Funktionen

Eine Menge radialer Funktionen

$$\mathcal{R} := \left\{ \Phi : Z \to \mathbb{R}, (\cdot, t) \mapsto \phi(\|\cdot - t\|_2), \quad \phi : \mathbb{R}_0^+ \to \mathbb{R} \text{ messbar}, t \in Z \subset \mathbb{R}^k \right\}$$

kann als Verknüpfung aus zwei Mengen $\mathcal{R} = \mathcal{M} \circ \mathcal{F}_{Ball}$ dargestellt werden:

$$\mathcal{F}_{Ball} = \{g : Z \to \mathbb{R}_0^+, z \mapsto \|z - t\|_2, \qquad t \in Z\},$$
$$\mathcal{M} = \{\phi : \mathbb{R}_0^+ \to \mathbb{R}, x \mapsto \phi(x), \qquad \phi \text{ messbar}\}.$$

Lemma

Falls alle Funktionen aus $\mathcal{M}$ monoton sind, gilt:

$$\mathrm{VCD}(\mathcal{M} \circ \mathcal{F}_{Ball}) \leq 2\,\mathrm{VCD}(\mathcal{F}_{Ball}) + 1$$

Monotoniebedingung

Mit dieser Idee setzen wir die Funktion ϕ zusammen:

$$\phi(z) = \sum_{i=1}^{p} \phi|_{\mathcal{I}_i}(z), \qquad \mathcal{I}_1 \cup \ldots \cup \mathcal{I}_p = \mathbb{R}_0^+, \quad \mathcal{I}_i \cap \mathcal{I}_j = \emptyset, \quad \phi \text{ auf } \mathcal{I}_i \text{ monoton}, \quad 1 \leq i, j \leq p. \tag{1}$$

Wir setzen

$$\mathcal{M}_p := \{\phi : \mathbb{R}_0^+ \to [A, B], \phi \text{ erfüllt (1) für } p \in \mathbb{N}\}$$

und erhalten:

Satz

$$\mathrm{VCD}(\mathcal{M}_p \circ \mathcal{F}_{Ball}) \leq c_p \left(2\,\mathrm{VCD}(\mathcal{F}_{Ball}) + 1\right), \quad c_p \in \mathcal{O}(p \log(p))$$

Mit $\mathrm{VCD}(\mathcal{F}_{Ball}) = k + 1$ folgt:

$$\mathrm{VCD}(\mathcal{M}_p \circ \mathcal{F}_{Ball}) \leq c_p(2k + 3).$$

Ergebnisse für radiale Funktionen

Notation

$$\mathcal{F} = \mathcal{M}_p \circ \mathcal{F}_{Ball} = \{\phi(\|\cdot - t\|_2) : Z \to [A, B], \quad \phi \in \mathcal{M}_p, t \in Z\}$$

Satz

Mit einer Wahrscheinlichkeit von mindestens $1 - \delta$ *findet man Punkte* $z_1, \ldots z_n \in Z$, *sodass gilt:*

$$\sup_{f \in \mathcal{F}} \left| R(f) - \widehat{R}(f) \right| \leq C(B - A) \frac{1}{\sqrt{n}} \max \left\{ \sqrt{c_p(2k + 3)}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}.$$

Zusammenfassung

Verwednung von Schranken aus der statistischen Lerntheorie:

$$P\left\{z_1,\ldots,z_n \in Z : \sup_{f\in\mathcal{F}}\left|E[f(z)] - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \leq \alpha_{n,\delta,\mathcal{F}}\right\} \geq 1-\delta, \quad \alpha_{n,\delta,\mathcal{F}} \overset{n\to\infty}{\longrightarrow} 0$$

Übertragung in die Approximationstheorie:

$$\sup_{f\in\mathcal{F}}\left|\int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)\right| \leq C(B-A)\frac{1}{\sqrt{n}}\sqrt{\mathrm{VCD}(\mathcal{F})}$$

Anwendung auf radiale Basisfunktionen $\mathcal{F} = \{\phi(\cdot - t) : Z \to \mathbb{R},\ \phi \in \mathcal{M}_p, t \in Z\}$:

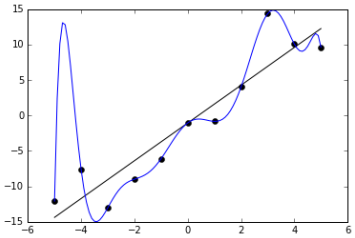
$$\mathrm{VCD}(\mathcal{F}) \leq c_p(2k+3)$$

Kernpunkte

$$\sup_{f\in\mathcal{F}}\left|\int_z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \leq \alpha_{\mathrm{VCD}(\mathcal{F}),[A,B]} \tag{1}$$

1. (1) gilt auf denselben Punkten $z_1, \ldots, z_n$ gleichmäßig für alle $f \in \mathcal{F}$
2. (1) ist nur abhängig von der VC-Dimension von $\mathcal{F}$, diese lässt sich speziell für radiale Funktionen leicht berechnen
3. Zusammenhang zwischen statistischer Lerntheorie und Approximationstheorie

Abbildungsverzeichnis



By Ghiles - Own work, CC BY-SA 4.0,
https://commons.wikimedia.org/w/index.php?curid=47471056

Literatur


Federico Girosi
Approximation Error Bounds That Use Vc-Bounds
Proc. Internat. Conf. Artificial Neural Networks 1995.


Giorgio Gnecco und Marcello Sanguineti
Approximation error bounds via Rademacher's complexity
Applied Mathematical Sciences, 2008.

Anwendung auf Span

$$\mathcal{S} := \left\{ \sum_{i=1}^{m} a_i f_i(z) : a_1, \ldots, a_m \in \mathbb{R} \right\}, \quad \mathcal{S}_M := \left\{ \sum_{i=1}^{m} a_i f_i(z) : \sum_{i=1}^{m} |a_i| \leq M, a_1, \ldots, a_m \in \mathbb{R} \right\}$$

Lemma

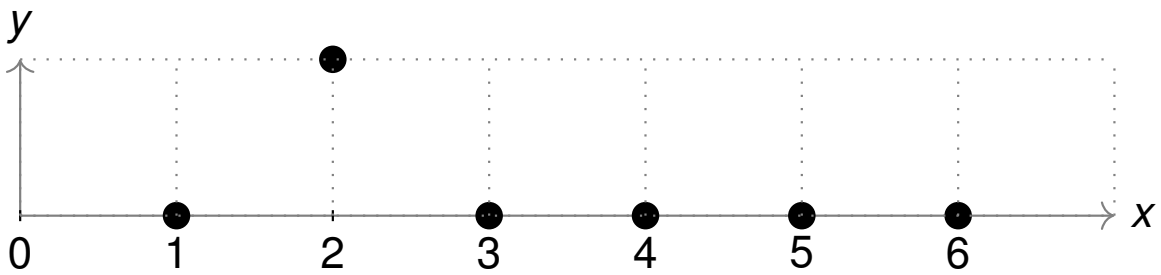
$$\sup_{f \in \mathcal{S}} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n} \sum_{i=1}^{n} f(\tilde{z}_i) \right| \leq (B-A)C\sqrt{\frac{m+1}{n}},$$

$$\sup_{f \in \mathcal{S}_M} \left| \int_Z f(z)\lambda(z)dz - \frac{1}{n} \sum_{i=1}^{n} f(z) \right| \leq (B-A)CM \max\{|A|, |B|\}\sqrt{\frac{2\ln(2m)}{n}}.$$

Beispiel: Indikatorfunktionen auf $\mathbb{N}$

$$\mathcal{F} := \{f_i(j) : \mathbb{N} \to \{0,1\}, f_i(j) = \delta_{i,j}, i \in \mathbb{N}\}.$$

Wir betrachten $f_2 \in \mathcal{F}$:



Wir berechnen die Wachstumsfunktion:

$$\mathcal{G}_{\mathcal{F}}(n) = n + 1.$$

Somit gilt:

$$\mathrm{VCD}(\mathcal{F}) = \max\{n \in \mathbb{N}_0 | \mathcal{G}_{\mathcal{F}}(n) = 2^n\} = 1.$$

Schranken: VCD und RK

Mit einer Wahrscheinlichkeit von mindestens $(1-\delta)$ gilt:

$$\sup_{f\in\mathcal{F}}\left|\int_Z f(z)\lambda(z)dz - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| \leq \alpha.$$

- Schranke mit der Wachstumsfunktion:

$$\alpha = (B-A)\sqrt{8\frac{\ln\left(\mathcal{G}_{\mathcal{F}}(n)\right) - \ln\left(\frac{\delta}{4}\right)}{n}}.$$

- Schranke mit der Rademacher Komplexität:

$$\alpha = (B-A)C\frac{1}{\sqrt{n}}\max\left\{\frac{\mathcal{R}_n(\mathcal{F}-A)}{B-A}, -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)}\right\}.$$

Dabei gilt:

$$\mathcal{G}_{\mathcal{F}}(n) \leq \left(\frac{en}{\mathrm{VCD}(\mathcal{F})}\right)^{\mathrm{VCD}(\mathcal{F})}$$

$$\mathcal{R}_n(\mathcal{F}-A) \leq C(B-A)\sqrt{\mathrm{VCD}(\mathcal{F})}.$$

Beweisübersicht für die VC-Schranke - Idee

Idee:

1. Nutze *Hoeffding*-Ungleichung mit $E[Z_i] = \int_Z Z_i(z)dP(z)$:

$$P\left(\left|\frac{1}{n}\sum_{i=1}^{n} Z_i(z) - E[Z_i]\right| \geq \epsilon\right) \leq 2\exp\left(-2n\epsilon^2(B-A)^{-2}\right).$$

2. Erweitere mit union-Bound $P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$ auf endlich viele Funktionen, dabei:

$$P(A_1 > \epsilon \cup A_2 > \epsilon) = P\left(\sup_{A \in \{A_1, A_2\}} A > \epsilon\right).$$

Beweisübersicht für die VC-Schranke - Umsetzung

Notation

$\mathcal{F}_{\mathcal{H}} = \{\mathcal{H}(f(\cdot - \beta)), f \in \mathcal{F}, \beta \in \mathbb{R}\}$

1. Es genügt, nur endlich viele Funktionen aus $\mathcal{F}$ zu betrachten:
   - Einschränkung auf Funktionen mit Bild in $\{0, 1\}$;
   - Einschränkung der Funktionsauswertungen auf $2n$ Punkte

   Damit betrachtet man höchstens 2^{2n} Funktionen:

$$P\left(\left\{\sup_{f\in\mathcal{F}}|R(f) - \widehat{R}(f)| > \epsilon\right\}\right) \leq P\left(\left\{\sup_{f\in\mathcal{F}_{\mathcal{H}}}\left|\int_Z f(z)dP(z) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right| > \frac{\varepsilon}{B-A}\right\}\right)$$
$$\leq 2P\left(\sup_{f\in\mathcal{F}_{\mathcal{H}}}\left|\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \frac{1}{n}\sum_{i=1}^{n} f(z_{i+n})\right| > \frac{\epsilon}{2(B-A)}\right)$$

2. Hoeffding-Ungleichung mit $E\left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - f(z_{i+n})\right] = 0$ und $[A, B] = [0, 1]$

Beweis zur Abschätzung mit der Rademacher Komplexität

Notation

$$D := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - E_P[f(x)]) \right|, \quad \mathcal{R}_n(\mathcal{F}) := E_P E_\epsilon \left[\sup_{f \in \mathcal{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^{n} \epsilon_i f(x_i) \right| \right]$$

1. Mit einer Wahrscheinlichkeit von mindestens $(1 - \delta)$ gilt:

$$nD \leq 2E_P[nD] + C\left(\sqrt{-n \ln(\delta)} + \ln(\delta)\right)$$

2. Symmetrisierung:

$$E_P[D] \leq 2\frac{\mathcal{R}_n(\mathcal{F})}{\sqrt{n}}.$$

3. Damit erhält man mit einer Wahrscheinlichkeit von mindestens $(1 - \delta)$:

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(z_i) - E_P[f(z)] \right| \leq \frac{4\mathcal{R}_n(\mathcal{F})}{\sqrt{n}} + \frac{\tilde{C}\sqrt{-\ln(\delta)}}{\sqrt{n}} - \frac{\tilde{C}}{n} \ln(\delta)$$

$$\leq (B - A)C \frac{1}{\sqrt{n}} \max\left\{ \mathcal{R}_n(\mathcal{F}), -\frac{\ln(\delta)}{\sqrt{n}}, \sqrt{-\ln(\delta)} \right\}.$$

Abschätzung der VCD gegen die Rademacher Komplexität

Notation

$\hat{\mathcal{R}}_S(\mathcal{F}) := E_\epsilon \left[\sup_{f\in\mathcal{F}} \frac{1}{\sqrt{n}} \left| \sum_{i=1}^n \epsilon_i f(x_i) \right| \right], \quad \mathcal{R}_n(\mathcal{F}) := E_P \left[\hat{\mathcal{R}}_S(\mathcal{F}) \right].$

Zu beweisen:

$$\mathcal{R}_n(\mathcal{F}) \leq (B-A) C \sqrt{\mathrm{VCD}(\mathcal{F})}.$$

1. Beweis über ε-Überdeckungszahlen:

$$\hat{\mathcal{R}}(\mathcal{F}) \leq \hat{C} \int_0^\infty (\ln \mathcal{N}(\varepsilon, \mathcal{F}, \mathcal{L}_2(\mu_n)))^{\frac{1}{2}} \, d\varepsilon$$

2. weiterhin gilt:

$$\mathcal{N}(\mathcal{F}, \mathcal{L}_2(\mu), \epsilon) \leq \left(\frac{2}{\varepsilon} \right)^{C_1 \left(\frac{2}{\varepsilon} - 1 \right) \mathrm{VCD}(\mathcal{F})}$$