'Empirical margin distributions and bounding the generalization error of combined classifiers' - Vladimir Koltchinskii & Dimitry Panchenko

Statsitcal Learning Seminar 4. Vortrag

Ben Deitmar

09.06.2020

Albert-Ludwigs-Universität Freiburg

Gliederung

- Einführung Empirische Prozesse
- Rademacher-Komplexität
- Lévi-Abstand
- Aussagen des Artikels
- Anwendungen
- Vergleich

Glivenko-Cantelli-Klassen

- P Verteilung auf $[0,1] \times \{-1,1\}$
- $\mathcal{F} \subset L^1(P)$
  d.h. $\mathcal{F} \subset \{f : [0,1] \times \{-1,1\} \to \mathbb{R}\ mb. \mid \mathbb{E}_P[|f(X,Y)|] < \infty\}$
- $((X_1,Y_1),...,(X_n,Y_n)) \sim P^{\otimes n}$: $\hat{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$

Definition:
$\mathcal{F}$ ist P-**Glivenko-Cantelli-Klasse** $\Leftrightarrow \sup_{f\in\mathcal{F}} \left|\hat{P}_n(f) - P(f)\right| \xrightarrow{n\to\infty}_{fs.} 0$

Glivenko-Cantelli-Klassen

- P Verteilung auf $[0,1] \times \{-1,1\}$
- $\mathcal{F} \subset L^1(P)$
  d.h. $\mathcal{F} \subset \{f : [0,1] \times \{-1,1\} \to \mathbb{R}\ mb. \mid \mathbb{E}_P[|f(X,Y)|] < \infty\}$
- $((X_1,Y_1),...,(X_n,Y_n)) \sim P^{\otimes n}$: $\hat{P}_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{(X_i,Y_i)}$

Definition:
$\mathcal{F}$ ist P-**Glivenko-Cantelli-Klasse** $\Leftrightarrow \sup_{f\in\mathcal{F}} \left|\hat{P}_n(f) - P(f)\right| \xrightarrow{n\to\infty}_{fs.} 0$

Problem: $\sup_{f\in\mathcal{F}} \left|\hat{P}_n(f) - P(f)\right|$ nicht immer messbar.
Kann man mit zusätzlichen Anforderungen an $\mathcal{F}$ beheben.
(siehe [3] Sektion 5.3)

Glivenko-Cantelli-Klassen

Starkes Gesetz der großen Zahlen:

$(X_1, Y_1), (X_2, Y_2), ... \overset{iv.}{\sim} P$

$\Rightarrow \hat{P}_n(f) - P(f) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, Y_i) - \mathbb{E}_P[f] \xrightarrow{n\to\infty}_{fs.} 0$

$\Rightarrow \{f\}$ ist P-GC.

Satz von Glivenko-Cantelli:

$\forall \widetilde{f} \in L^1(P): \sup_{z\in\mathbb{R}} |\hat{P}_n(\widetilde{f} \leq z) - P(\widetilde{f} \leq z)| \xrightarrow{n\to\infty}_{fs.} 0$

$\Rightarrow$ Für $\widetilde{f} \in L^1(P)$ ist $\{1_{\widetilde{f}\leq z} \mid z \in \mathbb{R}\}$ P-GC.

Donsker-Klassen

Definition:

$\mathcal{F} \subset L^2(P)$ ist P-**Donsker-Klasse**

$$\Leftrightarrow \exists G_P \text{ Prozess auf } \mathcal{F}: \sqrt{n}\left(\hat{P}_n(f) - P(f)\right)_{f\in\mathcal{F}} \xrightarrow{n\to\infty}_{\mathcal{L}} G_P$$

D.h. $\forall H \in C_b(\ell^\infty(\mathcal{F})):$

$$\mathbb{E}\left[H\left(\sqrt{n}\left(\hat{P}_n(f) - P(f)\right)_{f\in\mathcal{F}}\right)\right] \xrightarrow{n\to\infty} \mathbb{E}[H(G_P)]$$

Donsker-Klassen

Definition:
$\mathcal{F} \subset L^2(P)$ ist P-**Donsker-Klasse**

$$\Leftrightarrow \exists G_P \text{ Prozess auf } \mathcal{F}: \sqrt{n}\left(\hat{P}_n(f) - P(f)\right)_{f\in\mathcal{F}} \xrightarrow{n\to\infty}_{\mathcal{L}} G_P$$

D.h. $\forall H \in C_b(\ell^\infty(\mathcal{F})):$

$$\mathbb{E}\left[H\left(\sqrt{n}\left(\hat{P}_n(f) - P(f)\right)_{f\in\mathcal{F}}\right)\right] \xrightarrow{n\to\infty} \mathbb{E}[H(G_P)]$$

Zentraler Grenzwertsatz:
$\forall f \in L^2(P):$

$$\sqrt{n}\left(\hat{P}_n(f) - P(f)\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} f(X_i, Y_i) - \mathbb{E}_P[f] \xrightarrow{n\to\infty}_{\mathcal{L}} N(0, \sigma^2)$$

$\Rightarrow \{f\}$ ist P-Donsker

Donsker-Klassen (Eigenschaften)

- $\mathcal{F}$ P-Donsker und $\mathcal{G} \subset \mathcal{F} \Rightarrow \mathcal{G}$ P-Donsker
- $\mathcal{F}, \mathcal{G}$ P-Donsker $\Rightarrow \mathcal{F} \cup \mathcal{G}$ P-Donsker
- $\mathcal{F}$ P-Donsker $\Rightarrow co(\mathcal{F})$ P-Donsker
- $\mathcal{F}$ P-Donsker $\Rightarrow \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\hat{P}_n(f) - P(f)\right|\right] \leq \frac{C}{\sqrt{n}}$

Donsker-Klassen (Eigenschaften)

- $\mathcal{F}$ P-Donsker und $\mathcal{G} \subset \mathcal{F} \Rightarrow \mathcal{G}$ P-Donsker
- $\mathcal{F}, \mathcal{G}$ P-Donsker $\Rightarrow \mathcal{F} \cup \mathcal{G}$ P-Donsker
- $\mathcal{F}$ P-Donsker $\Rightarrow co(\mathcal{F})$ P-Donsker
- $\mathcal{F}$ P-Donsker $\Rightarrow \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left|\hat{P}_n(f) - P(f)\right|\right] \leq \frac{C}{\sqrt{n}}$

Universelle Donsker-Klassen (Dudley, 1992):
$\forall M > 0 \; \forall P$ Verteilung auf $\mathbb{R}$: E_M ist P-Donsker-Klasse
für $E_M = \{f : \mathbb{R} \to \mathbb{R} \mid M \geq ||f||_{TV}\}$.

(vgl. [3] Seite 329)

Lemma 1.4 Algorithmus-Risiko bei Donsker-Klassen

Für $y \in \{-1, 1\}$:

- $P_y(A) := P_{X|Y=y}(A) = P(X \in A \mid Y = y)$ Verteilung auf $[0, 1]$
- $L : \mathbb{R} \times \{-1, 1\} \to [0, \infty]$ Verlustfunktion
- Funktionenfamilie $(f_\alpha)_{\alpha \in \Lambda}$
- $\mathcal{F}_y := \{x \mapsto L(f_\alpha(x), y) \mid \alpha \in \Lambda\} \subset L^2(P_y)$
- falls $\mathcal{F}_y$ glm. beschränkt durch $M > 0$
- falls $\mathcal{F}_y$ P_y-Donsker-Klasse

dann existiert ein $C > 0$, sodass für die ERM gilt:

$$\mathbb{E}\left[R_{L,P}(\Phi(T_n)) - \inf_{\alpha \in \Lambda} R_{L,P}(f_\alpha)\right] \leq \frac{C}{\sqrt{n}}.$$

Lemma 1.4 (Beweis-Idee)

- $\mathcal{F}_y$ ist P_y-Donsker $\Rightarrow \mathbb{E}_{P_y^{\otimes n}}\left[\sup_{\alpha\in\Lambda}\left|\hat{P}_n(L(f_\alpha(\cdot),y)) - P(L(f_\alpha(\cdot),y))\right|\right] \leq \frac{c}{\sqrt{n}}$
- $\mathbb{E}\left[\left|\frac{n_y}{n} - P(Y=y)\right|\right]^2 \leq \mathbb{E}\left[\left(\frac{n_y}{n} - P(Y=y)\right)^2\right] = \mathbb{V}[\frac{n_y}{n}] = \frac{nP(Y=y)P(Y=-y)}{n^2} \leq \frac{1}{n}$
- $P(L(f_\alpha(X),Y)) = P(Y=1)P(L(f_\alpha(X),1)) + P(Y=-1)P(L(f_\alpha(X),-1))$

Lemma 1.4 (Beweis-Idee)

- $\mathcal{F}_y$ ist P_y-Donsker $\Rightarrow \mathbb{E}_{P_y^{\otimes n}}\left[\sup_{\alpha\in\Lambda}\left|\hat{P}_n(L(f_\alpha(\cdot),y)) - P(L(f_\alpha(\cdot),y))\right|\right] \leq \frac{c}{\sqrt{n}}$
- $\mathbb{E}\left[\left|\frac{n_y}{n} - P(Y=y)\right|\right]^2 \leq \mathbb{E}\left[\left(\frac{n_y}{n} - P(Y=y)\right)^2\right] = \mathbb{V}[\frac{n_y}{n}] = \frac{nP(Y=y)P(Y=-y)}{n^2} \leq \frac{1}{n}$
- $P(L(f_\alpha(X),Y)) = P(Y=1)P(L(f_\alpha(X),1)) + P(Y=-1)P(L(f_\alpha(X),-1))$

$$\mathbb{E}\left[\sup_{\alpha\in\Lambda}\left|\hat{P}_n(L(f_\alpha(\cdot),\cdot)) - P(L(f_\alpha(\cdot),\cdot))\right|\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[\sup_{\alpha\in\Lambda}\left|\hat{P}_n(L(f_\alpha(\cdot),Y_i)) - P(L(f_\alpha(\cdot),\cdot))\right| \mid Y_1,...,Y_n\right]\right]$$

$$\leq \mathbb{E}_{Y_i}\left[\sum_{y\in\{-1,1\}} \mathbb{E}_{P_y^{\otimes n_y}}\left[\sup_{\alpha\in\Lambda}\left|\frac{n_y}{n}\widehat{(P_y)}_{n_y}(L(f_\alpha(\cdot),y)) - P(Y=y)P_y(L(f_\alpha(\cdot)),y)\right|\right]\right]$$

$$\leq \mathbb{E}_{Y_i}\left[\sum_{y\in\{-1,1\}} \frac{n_y}{n}\frac{c}{\sqrt{n_y}} + M\left|\frac{n_y}{n} - P(Y=y)\right|\right] \leq \frac{2c}{\sqrt{n}} + 2M\frac{1}{\sqrt{n}}$$

□

Rademacher-Komplexität

- $\mathcal{F} \subset L^1(P)$ für P Verteilung auf $E \subset \mathbb{R}^d$
- $Z_1, Z_2, ...uiv. \sim P$
- $\varepsilon_1, \varepsilon_2, ...uiv. \sim U(\{-1, 1\})$
- $(\varepsilon_i)_{i \in \mathbb{N}}$ und $(Z_i)_{i \in \mathbb{N}}$ unabhängig

Definition:

$$\mathcal{RK}_n(\mathcal{F}) := \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i f(Z_i) \right| \right]$$

Bemerkung 2.2 : Symmetrisierung

Für jedes $\mathcal{F} \subset L^1(P)$ gilt:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\hat{P}_n(f) - P(f)\right|\right] \leq 2\mathcal{RK}_n(\mathcal{F}).$$

(Beweis in [4] Lemma 2.3.1)

Lemma 2.3 : Rademacher-Komplexität von Donsker-Klassen

Sei $\mathcal{F} \subset L^2(P)$ eine P-Donsker-Klasse, dann exisitert ein $C > 0$, sodass

$$\forall n \in \mathbb{N} : \ \mathcal{RK}_n(\mathcal{F}) \leq \frac{C}{\sqrt{n}}.$$

(Beweis in der Ausarbeitung)

Lemma 2.4 : Rademacher-Komplexität von GC.-Klassen

Sei $\mathcal{F} \subset L^1(P)$ eine P-GC.-Klasse, dann gilt

$$\mathcal{RK}_n(\mathcal{F}) \xrightarrow{n\to\infty} 0.$$

(Beweis in der Ausarbeitung)

Lévi-Abstand

- P_1, P_2 Verteilungen auf $(E, \mathcal{A})$
- $f : E \to \mathbb{R}$ messbar

Definition:

$$\mathbb{L}_f(P_1, P_2) := \inf \{\delta > 0 \mid \forall t \in \mathbb{R} : P_i(f \leq t) \leq P_j(f \leq t + \delta) + \delta \\ , \ i, j \in \{1, 2\}\}.$$

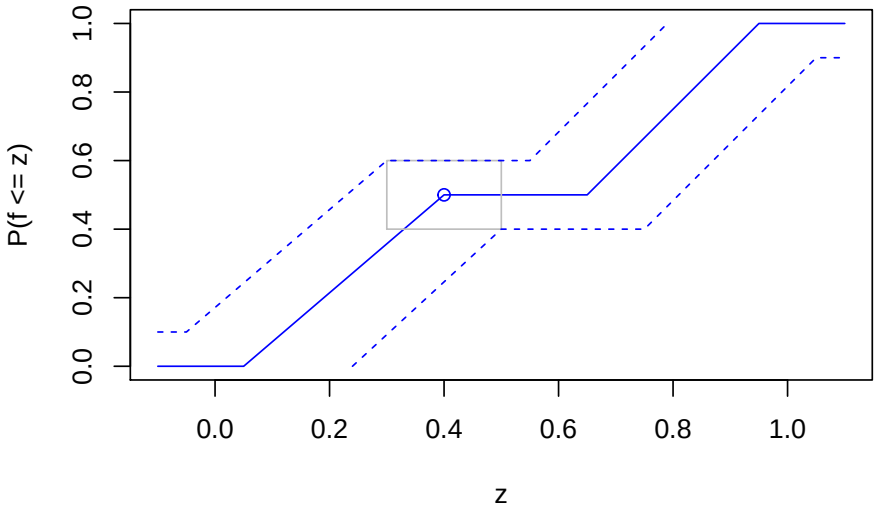
Lévi-Abstand

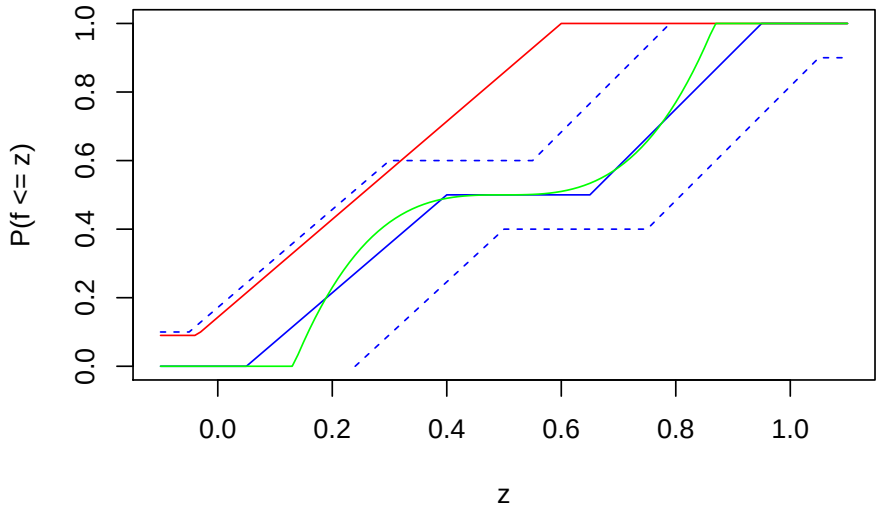
- P_1, P_2 Verteilungen auf $(E, \mathcal{A})$
- $f : E \to \mathbb{R}$ messbar

Definition:

$$\mathbb{L}_f(P_1, P_2) := \inf \{\delta > 0 \mid \forall t \in \mathbb{R} : P_i(f \leq t) \leq P_j(f \leq t + \delta) + \delta \\ , \ i, j \in \{1, 2\}\}.$$

Anschaulich: Abstand der Graphen von $F_{(f)_* P_1}$ und $F_{(f)_* P_2}$.





Eigenschaften Lévi-Abstand

- $P, P_1, P_2, \ldots$ Verteilungen auf $(E, \mathcal{A})$
- $f : E \to [0, 1]$ messbar

Dann gilt:

- $|P_1(f) - P_2(f)| \leq 4\,\mathbb{L}_f(P_1, P_2)$
- $(f)_* P_m \xrightarrow{m\to\infty}_{\mathcal{L}} (f)_* P \;\Leftrightarrow\; \mathbb{L}_f(P_m, P) \xrightarrow{m\to\infty} 0$

(mehr Eigenschaften **hier**)

Satz 3.1 : Abschätzung der Verteilungsfunktionen mit der Rademacherkomplexität

- $\mathcal{F} \subset \{f : E \to \mathbb{R} \text{ mb.}\}$ gleichmäßig beschränkt
- $(\varphi_k)_{k\in\mathbb{N}}$ Familie von Lipschitz-stetigen Funktionen mit $|\varphi_k(x) - \varphi_k(y)| \leq c_k |x - y|$ und $\varphi_k \geq 1_{(-\infty,0]}$

Dann gilt für jedes $t > 0$, dass

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \; P(f \leq 0) > \frac{t}{\sqrt{n}} + \inf_{k\in\mathbb{N}} \left[\hat{P}_n(\varphi_k \circ f) + 4c_k \mathcal{RK}_n(\mathcal{F}) + \left(\frac{\log(k)}{n} \right)^{\frac{1}{2}} \right] \right) \leq 2e^{-2t^2}$$

(Beweis-Idee in der Ausarbeitung)

Satz 3.2 : Abschätzung des Lévy-Abstandes

- $\mathcal{F} \subset \{f : E \to [-M, M] \text{ mb.}\}$

dann gilt für jedes $t > 0$, dass

$$\mathbb{P}\left(\sup_{f \in \mathcal{F}} \mathbb{L}_f(P, \hat{P}_n) \geq \frac{t}{\sqrt{n}} + 2\left(\frac{M}{\sqrt{n}} + \mathcal{RK}_n(\mathcal{F})\right)^{\frac{1}{2}}\right) \leq e^{-2t^2}$$

(Beweis-Idee in der Ausarbeitung)

Folgerungen

Für $\mathcal{F} \subset \{f : E \to [-M, M] \text{ mb.}\}$ gilt:

- $\mathcal{F}$ ist P-GC. $\Leftrightarrow \sup_{f \in \mathcal{F}} \mathbb{L}_f(P, \hat{P}_n) \xrightarrow{n \to \infty}_{fs.} 0$
- $\mathcal{F}$ ist P-Donsker $\Rightarrow \exists C > 0 \; \forall n \in \mathbb{N} : \; \mathbb{E}\left[\sup_{f \in \mathcal{F}} \mathbb{L}_f(P, \hat{P}_n)\right] \leq C \, n^{-\frac{1}{4}}$

(Beweise in der Ausarbeitung)

Folgerungen

Für $\mathcal{F} \subset \{f : E \to [-M, M] \text{ mb.}\}$ gilt:

- $\mathcal{F}$ ist P-GC. $\Leftrightarrow \sup_{f \in \mathcal{F}} \mathbb{L}_f(P, \hat{P}_n) \xrightarrow{n \to \infty}_{fs.} 0$
- $\mathcal{F}$ ist P-Donsker $\Rightarrow \exists C > 0 \; \forall n \in \mathbb{N} : \; \mathbb{E}\left[\sup_{f \in \mathcal{F}} \mathbb{L}_f(P, \hat{P}_n)\right] \leq C\, n^{-\frac{1}{4}}$

(Beweise in der Ausarbeitung)

Vergleich:

$\mathcal{F}$ ist P-Donsker $\Rightarrow \exists C > 0 \; \forall n \in \mathbb{N} : \; \mathbb{E}\left[\sup_{f \in \mathcal{F}} |P(f) - \hat{P}_n(f)|\right] \leq C\, n^{-\frac{1}{2}}$

Aussagen über Lévy-Abstand sind also nicht besonders nützlich für Risiko-Abschätzungen.

Satz 3.5 : Abschätzung mit Margin

- P Verteilung auf $E \times \{-1, 1\}$
- $\mathcal{F} \subset \{f : E \times \{-1, 1\} \to \mathbb{R}\, mb.\}$
  (Anschaulich: f ordnet x die Kategorie y zu, falls $f(x, y) > f(x, -y)$)
- Margin: $m_f(x, y) := f(x, y) - f(x, -y)$
- $RK_n := \mathcal{RK}_n(\{f(\cdot, 1) \mid f \in \mathcal{F}\} \cup \{f(\cdot, -1) \mid f \in \mathcal{F}\})$

Dann gilt für jedes $t > 0$:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \; P(m_f \leq 0) > \inf_{\delta \in (0,1)} \left[\hat{P}_n(m_f \leq \delta) + \frac{48}{\delta} RK_n + \sqrt{\frac{\log \log_2(\frac{2}{\delta})}{n}} \right] + \frac{t}{\sqrt{n}} \right) \leq 2e^{-2t^2}$$

Satz 3.5 : Abschätzung mit Margin

- P Verteilung auf $E \times \{-1, 1\}$
- $\mathcal{F} \subset \{f : E \times \{-1, 1\} \to \mathbb{R}\, mb.\}$
  (Anschaulich: f ordnet x die Kategorie y zu, falls $f(x, y) > f(x, -y)$)
- Margin: $m_f(x, y) := f(x, y) - f(x, -y)$
- $RK_n := \mathcal{RK}_n(\{f(\cdot, 1) \mid f \in \mathcal{F}\} \cup \{f(\cdot, -1) \mid f \in \mathcal{F}\})$

Dann gilt für jedes $t > 0$:

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \; P(m_f \leq 0) > \inf_{\delta \in (0,1)} \left[\hat{P}_n(m_f \leq \delta) + \frac{48}{\delta} RK_n + \sqrt{\frac{\log \log_2(\frac{2}{\delta})}{n}} \right] + \frac{t}{\sqrt{n}} \right) \leq 2e^{-2t^2}$$

$\Rightarrow$ Es lassen sich Aussagen über $R_P(\text{sign}(\Phi(T_n)))$ machen.

Anwendungen

- Artikel findet Zusammenhang zwischen Lévi-Abstand und Rademacher-Komplexität
- (Bem. 4.2) $\mathcal{RK}_n(\mathcal{F}) = \mathcal{RK}_n(co(\mathcal{F})) \Rightarrow$ interessant für z.B. 'Boosting'-Algorithmen
- Klassifikatoren $\text{sign} \circ \Phi(T_n)$ um Entscheidungsgrenze herum instabil, Artikel liefert Methoden damit umzugehen.

Satz 5.1 : Donsker-Klassen und RKHS

- P Verteilung auf $[0,1] \times \{-1,1\}$
- H_σ RKHS zu $K_\sigma(x,y) = e^{-\frac{(x-y)^2}{\sigma^2}}$
- $(f_\alpha)_{\alpha \in \Lambda_{r,\sigma}} = \mathcal{F}_{r,\sigma} := \{f \in H_\sigma \mid r \geq ||f||_{H_\sigma}\}$
- Verlustfunktion $L : \mathbb{R} \times \{-1,1\}$ im 1. Arg. Lipschitz-stetig
- ERM $\Phi_{r,\sigma} := \arg\min_{f \in \mathcal{F}_{r,\sigma}} R_{L,\hat{P}_n}(f)$

Dann gilt:

$$\exists C_1, C_2 > 0 \; \forall n \in \mathbb{N} \; \forall r, \sigma > 0 :$$
$$\mathbb{E}\left[R_{L,P}(\Phi_{r,\sigma}(T_n)) - \inf_{\alpha \in \Lambda_{r,\sigma}} R_{L,P}(f_\alpha)\right] \leq \frac{r}{\sigma}\frac{C_1}{\sqrt{n}} + \frac{C_2}{\sqrt{n}}$$

Satz 5.1 : Donsker-Klassen und RKHS (Beweis-Idee)

Letzter Vortrag: $\forall f \in H_\sigma : \ ||f||_{TV} \leq \sqrt{\frac{2}{\sigma^2}}||f||_{H_\sigma}$

$\Rightarrow \ \forall f \in \mathcal{F}_{r,\sigma} : \ ||f||_{TV} \leq \sqrt{\frac{2}{\sigma^2}}r$

$\Rightarrow \ \mathcal{F}_{r,\sigma} \subset E_{\frac{r}{\sigma}\sqrt{2}} := \{f : [0,1] \to \mathbb{R} \mid \frac{r}{\sigma}\sqrt{2} \geq ||f||_{TV}\}$

L Lipschitz $\overset{\Rightarrow}{}$ $\frac{\sigma}{r}L(\mathcal{F}_{r,\sigma}, -1) \cup \frac{\sigma}{r}L(\mathcal{F}_{r,\sigma}, 1) \subset E_C$

Satz 5.1 : Donsker-Klassen und RKHS (Beweis-Idee)

Letzter Vortrag: $\forall f \in H_\sigma : \ ||f||_{TV} \leq \sqrt{\frac{2}{\sigma^2}}||f||_{H_\sigma}$

$\Rightarrow \ \forall f \in \mathcal{F}_{r,\sigma} : \ ||f||_{TV} \leq \sqrt{\frac{2}{\sigma^2}} r$

$\Rightarrow \ \mathcal{F}_{r,\sigma} \subset E_{\frac{r}{\sigma}\sqrt{2}} := \{f : [0,1] \to \mathbb{R} \mid \frac{r}{\sigma}\sqrt{2} \geq ||f||_{TV}\}$

L Lipschitz $\Rightarrow \ \frac{\sigma}{r} L(\mathcal{F}_{r,\sigma}, -1) \cup \frac{\sigma}{r} L(\mathcal{F}_{r,\sigma}, 1) \subset E_C$

[3] Seite 329 $\Rightarrow$ E_C ist uniform Donsker.

$\Rightarrow \ \frac{\sigma}{r} L(\mathcal{F}_{r,\sigma}, i)$ sind $P_i := P_{X|Y=i}$-Donsker

$$\Rightarrow \ \exists c > 0 : \ \mathbb{E}\left[\sup_{f \in \mathcal{F}_{r,\sigma}} \left| P_i\left(\tfrac{\sigma}{r} L\left(f(X), i\right)\right) - \widehat{(P_i)}_n\left(\tfrac{\sigma}{r} L\left(f(X), i\right)\right)\right|\right] \leq \frac{c}{\sqrt{n}}$$

(Trick mit Bedingung auf Y aus Lemma 1.4)

$$\Rightarrow \ \mathbb{E}\left[\sup_{f \in \mathcal{F}_{r,\sigma}} \left| R_{L,P}(f) - \inf_{g \in \mathcal{F}_{r,\sigma}} R_{L,P}(g)\right|\right] \leq \frac{r}{\sigma}\frac{C_1}{\sqrt{n}} + \frac{C_2}{\sqrt{n}} \qquad \square$$

Satz 5.2 : RKHS. Familien-Risiko

- P Verteilung auf $[0,1] \times \{-1,1\}$
- $\mathcal{F}_{r,\sigma} := \{f \in H_\sigma \mid r \geq ||f||_{H_\sigma}\}$
- Verlustfunktion l Hinge-Loss
- P hat GNE $\alpha \in (0,\infty)$

Dann gilt:

$$\exists c, C_\alpha > 0 \; \forall n \in \mathbb{N} \; \forall r, \sigma > 0 :$$

$$\inf_{f \in \mathcal{F}_{r,\sigma}} R_{l,P}(f) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f) \leq c \left(\frac{1}{\sigma r^2} + C_\alpha 2^{\frac{\alpha}{2}} \sigma^\alpha \right)$$

(C_α wie in der Definition des GNE aus dem letzten Vortrag)
(Direkte Folgerung aus Thm. 2.7 in [2])

Bemerkung 5.4 : ERM in größer werdenden Familien

Wähle σ_n, r_n, sodass $\mathbb{E}\left[R_{l,P}(\Phi_{r_n,\sigma_n}) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f)\right] \searrow 0$

- 5.1 $\Rightarrow$ $\mathbb{E}\left[R_{l,P}(\Phi_{r_n,\sigma_n}(T_n)) - \inf_{f\in\mathcal{F}_{r_n,\sigma_n}} R_{l,P}(f)\right] \leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}} + \frac{c}{\sqrt{n}}$
- 5.2 $\Rightarrow$ $\inf_{f\in\mathcal{F}_{r_n,\sigma_n}} R_{l,P}(f) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f) \leq \frac{c}{\sigma_n r_n^2} + c\sigma_n^{\alpha}$

Bemerkung 5.4 : ERM in größer werdenden Familien

Wähle σ_n, r_n, sodass $\mathbb{E}\left[R_{l,P}(\Phi_{r_n,\sigma_n}) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f)\right] \searrow 0$

- 5.1 $\Rightarrow$ $\mathbb{E}\left[R_{l,P}(\Phi_{r_n,\sigma_n}(T_n)) - \inf_{f\in\mathcal{F}_{r_n,\sigma_n}} R_{l,P}(f)\right] \leq \frac{r_n}{\sigma_n}\frac{c}{\sqrt{n}} + \frac{c}{\sqrt{n}}$
- 5.2 $\Rightarrow$ $\inf_{f\in\mathcal{F}_{r_n,\sigma_n}} R_{l,P}(f) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f) \leq \frac{c}{\sigma_n r_n^2} + c\sigma_n^{\alpha}$

$\Rightarrow$ $\mathbb{E}\left[R_{l,P}(\Phi_{\lambda_n,\sigma_n}) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f)\right] \leq c\ \max\left(\frac{1}{\sigma_n r_n^2}, \sigma_n^{\alpha}, \frac{r_n}{\sigma_n\sqrt{n}}, \frac{1}{\sqrt{n}}\right)$

Wähle $\sigma_n = n^{-\frac{1}{3(\alpha+1)}}$ und $r_n = n^{\frac{1}{6}}$

$\Rightarrow$ $\mathbb{E}\left[R_{l,P}(\Phi_{r_n,\sigma_n}) - \inf_{f:[0,1]\to\{-1,1\}} R_{l,P}(f)\right] \leq c\, n^{-\frac{\alpha}{3(\alpha+1)}}$

Satz 5.5 : KDE mit Margin-Abschätzung

- P Verteilung auf $[0,1] \times \{-1,1\}$
- P habe TNE $q \in [0,\infty]$
- Voraussetzung:
  $\exists \varepsilon > 0 : \left| \int_{\mathbb{R}} \int_{\mathbb{R}} (p_y(x) - p_y(x + t\sigma)) \frac{e^{-t^2}}{\sqrt{2\pi}} \, dt \, dP_y(x) \right| \leq c \, \sigma^{\varepsilon}$
- $\sigma_n = n^{-\frac{1}{2(1+\varepsilon)}}$
- $\Phi_n^N(x) := \Phi^{Naiv}(T_n)(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma^2}}$

Satz 5.5 : KDE mit Margin-Abschätzung

- P Verteilung auf $[0,1] \times \{-1,1\}$
- P habe TNE $q \in [0,\infty]$
- Voraussetzung:
  $\exists \varepsilon > 0 : \left| \int_{\mathbb{R}} \int_{\mathbb{R}} (p_y(x) - p_y(x + t\sigma)) \frac{e^{-t^2}}{\sqrt{2\pi}} \, dt \, dP_y(x) \right| \leq c \, \sigma^{\varepsilon}$
- $\sigma_n = n^{-\frac{1}{2(1+\varepsilon)}}$
- $\Phi_n^N(x) := \Phi^{Naiv}(T_n)(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \frac{K_{\sigma_n}(x, X_i)}{\sqrt{2\pi\sigma^2}}$

Dann $\exists C > 0 \; \forall n \in \mathbb{N}$:

$$\mathbb{E}\left[R_P(\text{sign}(\Phi_n^N)) - \inf_{f:[0,1]\to\{-1,1\}} R_P(f) \right] \leq C \, n^{-\frac{\varepsilon q}{2(q+1)(1+\varepsilon)}}$$

(Beweis in der Ausarbeitung, verwende 3.5)

Satz 5.6 : ERMR mit Margin-Abschätzung

- P Verteilung auf $[0,1] \times \{-1,1\}$
- P habe GNE $\alpha \in (0,\infty)$
- $\lambda_n = n^{-\frac{1}{3}}$ und $\sigma_n = n^{-\frac{1}{3(\alpha+1)}}$
- ERMR $\Phi_{\lambda_n,\sigma_n}(T_n) = \arg\min_{f \in H_{\sigma_n}} \left(\lambda_n ||f||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(f)\right)$

  letzter Vortrag $\Rightarrow$ $\Phi_{\lambda_n,\sigma_n}(T_n)(x) = \sum_{i=1}^{n} \alpha_i^n Y_i K_{\sigma_n}(x, X_i)$
- Voraussetzung: $\exists \varepsilon \in [0, \frac{1}{2}) : \ \mathbb{E}\left[\sum_{j=1}^{n} \alpha_j^n\right] = \mathcal{O}(n^{\varepsilon})$

Satz 5.6 : ERMR mit Margin-Abschätzung

- P Verteilung auf $[0,1] \times \{-1,1\}$
- P habe GNE $\alpha \in (0, \infty)$
- $\lambda_n = n^{-\frac{1}{3}}$ und $\sigma_n = n^{-\frac{1}{3(\alpha+1)}}$
- ERMR $\Phi_{\lambda_n,\sigma_n}(T_n) = \arg\min_{f \in H_{\sigma_n}} \left(\lambda_n ||f||^2_{H_{\sigma_n}} + R_{L,\hat{P}_n}(f)\right)$

  letzter Vortrag $\Rightarrow$ $\Phi_{\lambda_n,\sigma_n}(T_n)(x) = \sum_{i=1}^{n} \alpha_i^n Y_i K_{\sigma_n}(x, X_i)$

- Voraussetzung: $\exists \varepsilon \in [0, \frac{1}{2}) : \ \mathbb{E}\left[\sum_{j=1}^{n} \alpha_j^n\right] = \mathcal{O}(n^{\varepsilon})$

Dann $\exists C > 0 \ \forall n \in \mathbb{N}$:

$$\mathbb{E}\left[R_P(\text{sign}(\Phi_{\lambda_n,\sigma_n}(T_n))) - \inf_{f:[0,1]\to\{-1,1\}} R_P(f)\right] \leq C\, n^{-\min\left(\frac{1-2\varepsilon}{4}, \frac{\alpha}{3(\alpha+1)}\right)}$$

(Beweis in der Ausarbeitung, verwende 3.5)

Quellen und Verweise

[1] V. Koltchinskii and D. Panchenko, *Empirical margin distributions and bounding the generalization error of combined classifiers*, Ann. Statist. **30** (2002), no. 1, 1–50, DOI 10.1214/aos/1015362182.

[2] Ingo Steinwart and Clint Scovel, *Fast rates for support vector machines using Gaussian kernels*, Ann. Statist. **35** (2007), no. 2, 575–607, DOI 10.1214/009053606000001226.

[3] R. M. Dudley, *Uniform central limit theorems*, 2nd ed., Cambridge Studies in Advanced Mathematics, vol. 142, Cambridge University Press, New York, 2014.

[4] Aad W. van der Vaart and Jon A. Wellner, *Weak convergence and empirical processes*, Springer Series in Statistics, Springer-Verlag, New York, 1996. With applications to statistics.

Vielen Dank für Ihre Aufmerksamkeit!