Risk Bounds of Statistical Learning

Abschätzungen des Erwarteten Risikos verschiedener Modelle auf Grundlage von Vapnik und Chervonenkis

Bastian Schnitzer

Universität Freiburg

20. Juli 2020

Inhalt

Inhalt

Grundlegendes

Inhalt

Grundlegendes

- Erinnerung

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Inhalt

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Motivation

- Resultat von Vapnik und Chervonenkis

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Motivation

- Resultat von Vapnik und Chervonenkis
- Bewertung dieser Abschätzung

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Motivation

- Resultat von Vapnik und Chervonenkis
- Bewertung dieser Abschätzung

Hauptresultate

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Motivation

- Resultat von Vapnik und Chervonenkis
- Bewertung dieser Abschätzung

Hauptresultate

- Abschätzung nach oben

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Motivation

- Resultat von Vapnik und Chervonenkis
- Bewertung dieser Abschätzung

Hauptresultate

- Abschätzung nach oben
- Abschätzungen nach unten

Inhalt

Grundlegendes

- Erinnerung
- VC-Klassen
- Entropie mit Klammern

Motivation

- Resultat von Vapnik und Chervonenkis
- Bewertung dieser Abschätzung

Hauptresultate

- Abschätzung nach oben
- Abschätzungen nach unten

Beweise (Ideen)

Klassifikationsproblem

- (X, Y) Paar von Zufallsvariablen auf $\mathcal{X} \times \{0, 1\}$
- P die zugehörige gemeinsame (unbekannte) Verteilung
- Trainingsmenge $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid nach P
- Klassifikator $t : \mathcal{X} \to \{0, 1\}$
- Wahrscheinlichkeit für Missklassifikation $P(Y \neq t(X))$

Bayes-Klassifikator

- Wahrscheinlichkeit für Missklassifikation $P(Y \neq t(X))$
- Regressions-Funktion $\eta(x) = P[Y = 1|X = x]$
- Bayes-Klassifikator $s^*(x) = \mathbb{1}_{\eta(x) \geq 1/2}$
- Wissen: (Vortrag 2) s^* minimiert Missklasifikation
- Wissen auch: $\eta(x)$ nahe bei $\frac{1}{2}$ ist schlecht

Loss-Funktion

- Bayes-Klassifikator s^* minimiert Missklasifikation
- Relative Loss-Funktion $\ell(s^*, t) = P(Y \neq t(X)) - P(Y \neq s^*(X))$ misst Genauigkeit eines Klassifikators t
- Erwartetes Risiko $\mathbb{E}[\ell(s^*, \hat{s})]$ zur Analyse eines Schätzers $\hat{s}$

Empirical Risk Minimization (ERM)

- System messbarer Mengen $\mathcal{A}$
- Modell $S = \{\mathbb{1}_A, A \in \mathcal{A}\}$ die Menge der zugehörigen Klassifikatoren
- Als Schätzer für s^* nehmen wir Minimierer des empirischen Kriteriums

$$t \to \gamma_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq t(X_i)}$$

- Ziel ist es, S so zu wählen, dass der Bias $\inf_{t \in S} \ell(s^*, t)$ klein genug bleibt, während
- S nicht zu "groß"wird

Regressionsmodell

- Betrachten $\xi_1, \ldots, \xi_n$ iid auf Raum $\mathcal{Z}$ mit gemeinsamer Verteilung P
- $\xi_i = (X_i, Y_i)$ iid nach P verteilt
- X Werte in $\mathcal{X}$
- Y Werte in $[0, 1]$
- Regressions-Funktion $\eta(x) = \mathbb{E}[Y|X = x]$
- Suchen Schätzer für η

Erwarteter Verlust

- Suchen Schätzer für $\eta = \mathbb{E}[Y|X = x]$
- $\mathcal{S}$ Menge aller messbaren Funktionen von $\mathcal{X}$ nach [0, 1]
- Loss-Funktion $\gamma : \mathcal{S} \times \mathcal{Z} \to [0, 1]$, so dass
- Gute Funktion $s = \arg\min_{t \in \mathcal{S}} \mathbb{E}[\gamma(t, \cdot)]$ minimiert den erwarteten Verlust
- Sprich: relativer Verlust $\ell(s, t) = \mathbb{E}[\gamma(t, \cdot) - \gamma(s, \cdot)] \geq 0 \quad \forall t \in \mathcal{S}$
- Wir wählen $\gamma(t, (x, y)) = (y - t(x))^2$, da η (bzw. s^*) das Minimum von $\mathbb{E}[(Y - t(X))^2]$ auf $\mathcal{S}$ liefern

Zerschmettern

Zerschmettern

Definition

Sei $\mathcal{A} \subseteq \mathcal{P}^X$ ein Mengensystem und $C \subseteq X$
Dann zerschmettert $\mathcal{A}$ die Menge C
$:\Leftrightarrow \{A \cap C : A \in \mathcal{A}\} = 2^C \Leftrightarrow |\{A \cap C : A \in \mathcal{A}\}| = 2^{|C|}$

- Sprich: wir können jede beliebige Teilmenge von C durch den Schnitt eines Elements von $\mathcal{A}$ mit C gewinnen

Beispiel

$X = \mathbb{R}^2, \mathcal{A} = \{\text{alle abgeschlossenen Halbebenen im } \mathbb{R}^2\}$,
$C = \{x_1, x_2, x_3 \in \mathbb{R}^2\}$

- $\mathcal{A}$ zerschmettert $\Leftrightarrow x_1, x_2, x_3$ nicht kolinear

VC-Klassen

Definition

Die VC-Dimension V eines Mengensystems $\mathcal{A}$ ist
$V = sup\{\#C : \mathcal{A} \text{ zerschmettert C}\}$

- Sprich: die größte Kardinalität von Mengen C, die von $\mathcal{A}$ zerschmettert werden
- Wenn beliebig große Mengen von $\mathcal{A}$ zerschmettert werden ist V wie erwartet ∞

Definition

Im Falle, dass $V < \infty$ nennen wir die Klasse $\mathcal{A}$ VC-Klasse

- Benannt nach Vladimir Vapnik und Alexey Chervonenkis

Entropie mit Klammern

Definition

Gegeben $f, g : \mathcal{X} \to \mathbb{R}$ bezeichnet $[f, g]$ die Menge der Funktionen $u : \mathcal{X} \to \mathbb{R}$ mit $f \leq u \leq g$

Definition

Eine ϵ-Klammer bezüglich $|| \cdot ||_{\mathbb{L}_1(\mu)}$ ist eine Klammer $[f, g]$ mit $||f - g||_{\mathbb{L}_1(\mu)} \leq \epsilon$

Definition

Die $\mathbb{L}_1(\mu)$-Entropie $H_1(\epsilon, S, \mu)$ bezeichnet den Logarithmus der kleinsten Anzahl von ϵ-Klammern $[f, g]$ bzgl $|| \cdot ||_{\mathbb{L}_1(\mu)}$, so dass diese S abdecken

ϵ-Netze

Definition

Sei (S, d) ein metrischer Raum. Wir nennen $G_\epsilon \subseteq \bar{S}$ ein ϵ-Netz, falls
$\forall \tilde{\theta} \in S \quad \exists \theta_0 \in G_\epsilon : d(\tilde{\theta}, \theta_0) \leq \epsilon$

Kullback-Leibler und Hellinger

Definition

Seien P, Q Wahrscheinlichkeitsmaße, die absolut-stetig bezüglich eines Maßes λ seien. Wir definieren den quadratischen Hellinger-Abstand als

$$\mathcal{H}^2(P,Q) = \frac{1}{2}\int \left(\sqrt{\frac{dP}{d\lambda}} - \sqrt{\frac{dQ}{d\lambda}}\right)^2 d\lambda$$

Definition

Für diskrete Wahrscheinlichkeitsmaße P, Q auf dem selben Raum $\mathcal{X}$ definieren wir die Kullback-Leibler-Information

$$\mathcal{K}(P,Q) = \sum_{x\in\mathcal{X}} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

Resultat von Vapnik-Chervonenkis

- $\mathcal{A}$ eine VC-Klasse, S das zugehörige Modell und VC-Dimension V
- $\mathcal{P}(S)$ die Menge aller gemeinsamen Verteilungen P, so dass $s^* \in S$
- (man beachte, dass sowohl η als auch s^* von P abhängen)
- $\hat{s}$ der Empirical Risk Minimizer (ERM)

Die Abschätzungen

- erhalten folgende Abschätzung für globale Konstanten κ_1 und κ_2 (Vapnik & Chervonenkis)

$$\sup_{P \in \mathcal{P}(S)} \mathbb{E}[\ell(s^*, \hat{s})] \leq \kappa_1 \sqrt{\frac{V}{n}}$$

und, falls $2 \leq V \leq n$:

$$\inf_{\tilde{s}} \sup_{P \in \mathcal{P}(S)} \mathbb{E}[\ell(s^*, \tilde{s})] \geq \kappa_2 \sqrt{\frac{V}{n}}$$

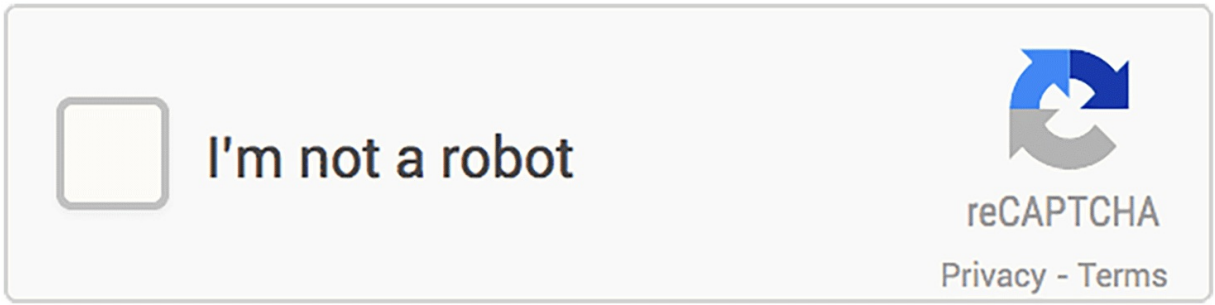
wobei das Infimum über alle Schätzer genommen wird

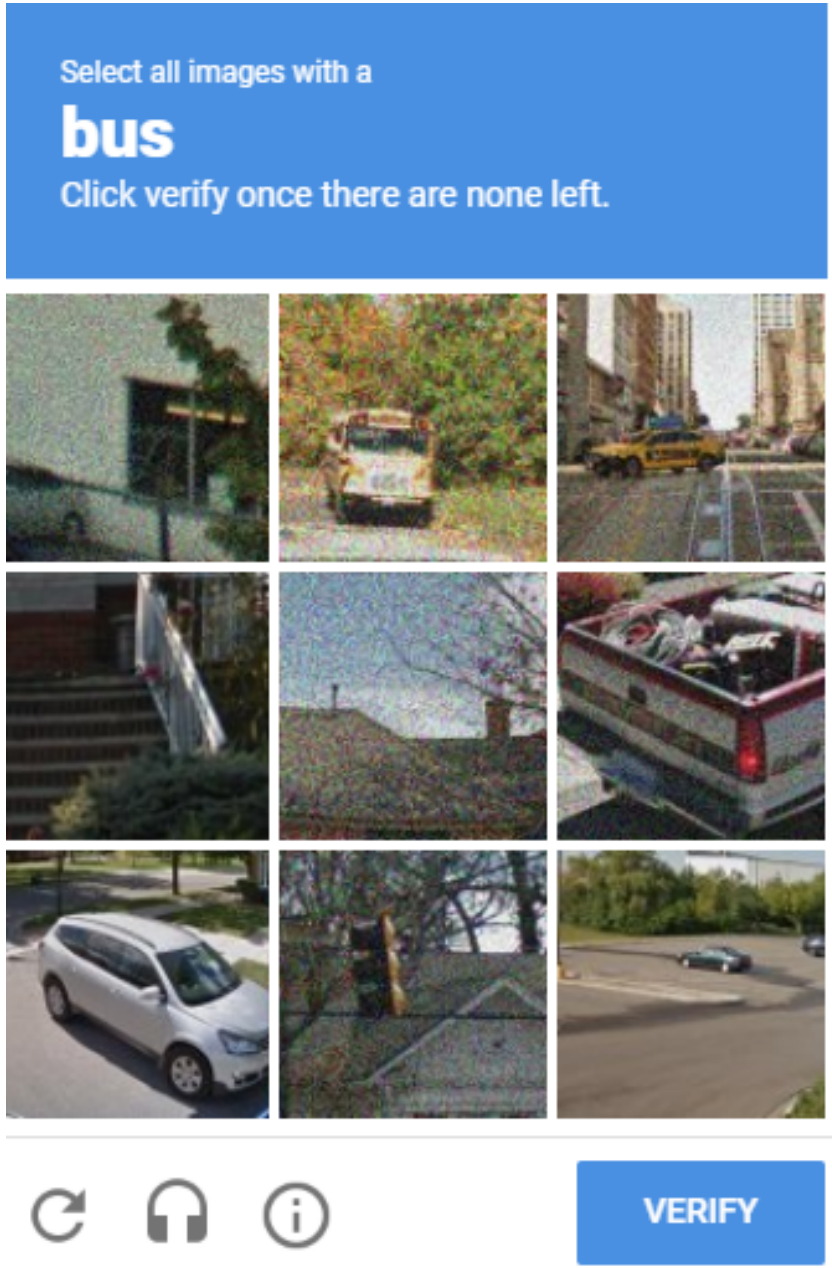
Beurteilung der Abschätzung von Vapnik Chervonenkis

- Diese Abschätzungen sind über-pessimistisch (also schwach)
- Man betrachte die über-optimistische Situation, wo $Y = \eta(X)$ (wir nennen P in diesem Fall zero-error Verteilung)
- Erhalten eine neue Minimax-Abschätzung

$$\inf_{\tilde{s}} \sup_{P\ zero-error} \mathbb{E}[\ell(s^*, \tilde{s})] \geq \kappa_2 \frac{V}{n}$$

- Das gibt Möglichkeiten für eventuell schärfere Abschätzungen, von denen wir einige in den Hauptresultate kennen lernen werden


I'm not a robot
reCAPTCHA
Privacy - Terms


Select all images with a
bus
Click verify once there are none left.
VERIFY

Vorbereitende Definitionen

Definition

Wir bezeichnen mit $\mathcal{C}_1$ die Klasse der monoton steigenden, stetigen Funktionen $\phi : \mathbb{R}_+ \to \mathbb{R}_+$, so dass $x \to \phi(x)/x$ monoton fallend auf $(0, +\infty)$ und $\phi(1) \geq 1$

Eigenschaft

Wir bezeichnen eine Menge S mit der Eigenschaft (M), wenn $\exists S' \subseteq S$ abzählbar s.d. $\forall t \in S$ gibt es eine Folge $(t_k) \subseteq S'$ s.d. $\forall \chi \in \mathcal{Z}, \gamma(t_k, \chi) \to \gamma(t, \chi), k \to \infty$

Definition

Wir bezeichnen mit d eine Halb-Metrik auf $\mathcal{S} \times \mathcal{S}$, so dass $Var_P[\gamma(t, \cdot) - \gamma(s, \cdot))] \leq d^2(s, t) \quad \forall t \in \mathcal{S}$

Weitere Begriffserklärungen

- Bias $\ell(s, S) = \inf_{t \in S} \ell(s, t)$
- Loss-Funktion $\gamma(t, (x, y)) = (y - t(x))^2$
- Relativer Loss $\ell(s, t) = \mathbb{E}[\gamma(t, \cdot) - \gamma(s, \cdot)] \quad \forall t \in \mathcal{S}$
- Empirische Loss-Funktion $\gamma_n(t) = \frac{1}{n} \sum_{i=1}^{n} \gamma(t, \xi_i)$
- Zentrierter empirischer Prozess $\bar{\gamma}_n(t) = \gamma_n(t) - \mathbb{E}[\gamma(t, \cdot)]$
- Die gute Funktion $s = \arg\min_{t \in \mathcal{S}} \mathbb{E}[\gamma(t, \cdot)]$
- Für $\rho \geq 0$ ist ein ρ-Schätzer $\hat{s}$ für s, so dass $\gamma_n(\hat{s}) \leq \rho + \inf_{t \in S} \gamma_n(t)$

Theorem 2

- Seien $\phi, \omega \in \mathcal{C}_1$
- Unser Modell $S \subseteq \mathcal{S}$ erfülle die Bedingung (M)
- Es gelte $d(s,t) \leq \omega(\sqrt{\ell(s,t)}) \quad \forall t \in \mathcal{S}$
- $\forall u \in S', \forall \sigma > 0 : \phi(\sigma) \leq \sqrt{n}\sigma^2$:

$$\sqrt{n}\mathbb{E}\left[\sup_{t \in S', d(u,t) \leq \sigma} [\bar{\gamma}_n(u) - \bar{\gamma}_n(t)]\right] \leq \phi(\sigma)$$

- ϵ_* sei die eindeutige Lösung von $\sqrt{n}\epsilon_*^2 = \phi(\omega(\epsilon_*))$

Dann exisiert eine globale Konstante κ, so dass

$$\forall y \geq 1 : \mathbb{P}[\ell(s, \hat{s}) \geq 2\rho + 2\ell(s, S) + \kappa\epsilon_*^2] \leq e^{-y}$$

Inbesondere gilt:

$$\mathbb{E}[\ell(s, \hat{s})] \leq 2(\rho + \ell(s, S) + \kappa\epsilon_*^2)$$

Talagrand's Ungleichung (Version von Bousquet)

- $\mathcal{F}$ abzählbare Familie messbarer Mengen
- $v, b > 0$ so dass
- $\forall f \in \mathcal{F} : ||f||_\infty \leq b, \sup_{f \in \mathcal{F}} Var(f) \leq v$
- $Z = sup_{f \in \mathcal{F}}(\gamma_n(f) - \gamma(f))$

Dann gilf für alle $y > 0$:

$$P\left[Z \geq E[Z] + \sqrt{2\frac{(v + 4bE[Z])y}{n}} + \frac{by}{n}\right] \leq e^{-y}$$

Lemma 5

- Seien S eine abzählbare Menge und $u \in S$
- $a : S \to \mathbb{R}_+$, so dass $a(u) = \inf_{t \in S} a(t)$
- Z ein von S indizierter Prozess
- Wir definieren $\mathcal{B}(\epsilon) = \{t \in S | a(t) \leq \epsilon\}$
- Es gelte $\sup_{t \in \mathcal{B}(\epsilon)} |Z(u) - Z(t)|$ habe endlichen Erwartungswert
- Sei $\psi : \mathbb{R}_+ \to \mathbb{R}_+$, so dass $\psi(x)/x$ monoton fallend
- Und erfülle $\mathbb{E}\left[\sup_{t \in \mathbb{B}(\epsilon)} [Z(u) - Z(t)]\right] \leq \psi(\epsilon)$ für alle $\epsilon \geq \epsilon_* > 0$

Dann gilt für alle $x \geq \epsilon_*$:

$$\mathbb{E}\left[\sup_{t \in S}\left[\frac{Z(u) - Z(t)}{a^2(t) + x^2}\right]\right] \leq 4x^{-2}\psi(x)$$

Beweis Theorem 2

- S erfüllt die Bedingung (M)

Finden also Folge $(t_k) \subseteq S', S'$ abzählbar, so dass für alle $t \in S$:

$$\mathbb{E}[\gamma(t_k, \cdot)] \to \mathbb{E}[\gamma(t, \cdot)]$$

Daraus folgern wir

$$\ell(s, S) = \ell(s, S')$$

Und finden einen Punkt $\pi(s) \in S'$ so dass gilt:

$$\ell(s, \pi(s)) \leq \ell(s, S) + \epsilon_*^2$$

Beweis Theorem 2

- Da $\hat{s}$ ein ρ-Schätzer, gilt: $\gamma_n(\hat{s}) \leq \rho + \inf_{t \in S} \gamma_n(t)$
- Definieren $x := \sqrt{\kappa' y}\epsilon_*$ für ein $\kappa' \geq 1$, das wir später wählen
- Definieren weiter (vielleicht hier der Einfachheit halber t aus S')

$$V_x := \sup_{t \in S} \frac{\bar{\gamma}_n(\pi(s)) - \bar{\gamma}_n(t)}{\ell(s,t) + \epsilon_*^2 + x^2}$$

Mit der Forderung, dass $V_x < 1/2$ erhalten wir folgende Abschätzung

$$\ell(s,\hat{s}) < 2(\rho + \ell(s,\pi(s))) + \epsilon_*^2 + x^2)$$

Und damit

$$\mathbb{P}[\ell(s,\hat{s}) \geq 2(\rho + \ell(s,S)) + 3\epsilon_*^2 + x^2] \leq \mathbb{P}[V_x \geq \tfrac{1}{2}]$$

Beweis Theorem 2

- Wir haben $\ell(s, \pi(s)) \leq \ell(s, S) + \epsilon_*^2$
- Da $Var_P[\gamma(t, \cdot) - \gamma(s, \cdot)] \leq d^2(s, t)$ für alle $t \in \mathcal{S}$
- Und $d(s, t) \leq \omega(\sqrt{\ell(s, t)})$ für alle $t \in \mathcal{S}$

Woraus wir folgern, dass

$$(Var_P[\gamma(t, \cdot) - \gamma(\pi(s), \cdot)])^{1/2} \leq 2\omega(\sqrt{\ell(s, t) + \epsilon_*^2})$$

Beweis Theorem 2

- Mit $(Var_P[\gamma(t,\cdot) - \gamma(\pi(s),\cdot)])^{1/2} \leq 2\omega(\sqrt{\ell(s,t) + \epsilon_*^2})$
- Und wir definieren $\omega_1 = 1 \wedge 2\omega$

Erhalten wir mithilfe der Monotonie-Bedingungen an ω, dass

$$\sup_{t \in S} Var_P \left[\frac{\gamma(t,\cdot) - \gamma(\pi(s),\cdot)}{\ell(s,t) + x^2} \right] \leq \frac{\omega_1^2(x)}{x^4}$$

Und auf der anderen Seite, da γ nur Werte in $[0,1]$ annimmt:

$$\sup_{t \in S} \left\| \frac{\gamma(t,\cdot) - \gamma(\pi(s),\cdot)}{\ell(s,t) + x^2} \right\|_\infty \leq \frac{1}{x^2}$$

Beweis Theorem 2

- Es gilt $\sup_{t \in S} Var_P[\frac{\gamma(t,\cdot)-\gamma(\pi(s),\cdot)}{\ell(s,t)+x^2}] \leq \frac{\omega_1^2(x)}{x^4}$
- Und $\sup_{t \in S} \| \frac{\gamma(t,\cdot)-\gamma(\pi(s),\cdot)}{\ell(s,t)+x^2} \|_\infty \leq \frac{1}{x^2}$

Wir verwenden Bousquet's Version von Talagrand's Ungleichung mit $v = \omega_1^2(x)x^{-4}$ und $b = x^{-2}$ und erhalten auf einer Menge Ω_y mit Wahrscheinlichkeit größer als $1 - e^{-y}$

$$V_x < \mathbb{E}[V_x] + \sqrt{\frac{2(\omega_1^2(x)x^{-2} + 4\mathbb{E}[V_x])y}{nx^2}} + \frac{y}{nx^2}$$

Beweis Theorem 2

- Weitere Schritte führen uns zur Erkenntnis, dass für alle $\epsilon \geq \epsilon_*^2$

$$\mathbb{E}\left[\sup_{t \in S', \ell(s,t) \leq \epsilon} (\bar{\gamma}_n(\pi(s)) - \bar{\gamma}_n(t))\right] \leq \phi(2\sqrt{2}\omega(\epsilon))$$

- Außerdem ist $\theta \to \phi(2\sqrt{2}\omega(\theta))/\theta$ monoton fallend

Wir verwenden Lemma 5, um zu zeigen, dass:

$$\mathbb{E}[V_x] \leq \frac{8\sqrt{2}}{\sqrt{\kappa'}}$$

Und Überlegungen an die Monotoniebedingungen von ω und ϕ bringen uns zu:

$$\frac{\omega_1^2(\epsilon_*^2)}{x^2} \leq 4n\epsilon_*^2$$

Beweis Theorem 2

- Haben $V_x < \mathbb{E}[V_x] + \sqrt{\frac{2(\omega_1^2(x)x^{-2}+4\mathbb{E}[V_x])y}{nx^2}} + \frac{y}{nx^2}$ auf Ω_y
- Und $\mathbb{E}[V_x] \leq \frac{8\sqrt{2}}{\sqrt{\kappa'}}$
- Mit weiteren Monotonie-Überlegungen und
- Der Definition von $x^2 = \kappa' y \epsilon_*^2$

Erhalten wir auf Ω_y, dass

$$V_x < \frac{8\sqrt{2}}{\sqrt{\kappa'}} + \sqrt{\frac{8(1+4/\sqrt{\kappa'})}{\kappa'}} + \frac{2}{3\kappa'}$$

Beweis Theorem 2

- Haben $\mathbb{P}[\ell(s,\hat{s}) \geq 2(\rho + \ell(s,S)) + 3\epsilon_*^2 + x^2] \leq \mathbb{P}[V_x \geq \frac{1}{2}]$
- Und $V_x < \frac{8\sqrt{2}}{\sqrt{\kappa'}} + \sqrt{\frac{8(1+4/\sqrt{\kappa'})}{\kappa'}} + \frac{2}{3\kappa'}$ auf Ω_y, wobei $\mathbb{P}(\Omega_y) > 1 - e^{-y}$

Wir können nun also κ' so wählen, dass $V_x < 1/2$ und erhalten:

$$\mathbb{P}[\ell(s,\hat{s}) \geq 2(\rho + \ell(s,S)) + x^2 + 3\epsilon_*^2] \leq \mathbb{P}(\Omega_y^c) \leq e^{-y}$$

Und für $\kappa = \kappa' + 3$ erhalten wir die gewünschte Abschätzung:

$$\mathbb{P}[\ell(s,\hat{s}) \geq 2\rho + 2\ell(s,S) + \kappa\epsilon_*^2] \leq e^{-y}$$

Integration liefert:

$$\mathbb{E}[\ell(s,\hat{s})] \leq 2(\rho + \ell(s,S) + \kappa\epsilon_*^2)$$

Begriffserklärung

- Für alle $h \in [0,1]$ bezeichne $\mathcal{P}(h,S)$ die Menge aller gemeinsamen Verteilungen P mit $s^* \in S$ und $|2\eta(x)-1| \geq h \quad \forall x \in \mathcal{X}$
- Und $\mathcal{P}(h,S,\mu)$ die Menge der Verteilungen P in $\mathcal{P}(h,S)$ mit vorgeschriebener Randverteilung μ auf $\mathcal{X}$
- Wir betrachten im Folgenden das Minimax Risiko

$$R_n(h,S) = \inf_{\hat{s}\in S} \sup_{P\in\mathcal{P}(h,S)} \mathbb{E}[\ell(s^*,\hat{s})]$$

- Und entsprechend

$$R_n(h,S,\mu) = \inf_{\hat{s}\in S} \sup_{P\in\mathcal{P}(h,S,\mu)} \mathbb{E}[\ell(s^*,\hat{s})]$$

- Wobei das Infimum jeweils über alle Schätzer auf Grundlage von n Testdaten genommen wird

Theorem 4

- Gegeben $h \in [0, 1]$
- Falls $\mathcal{A}$ VC-Klasse mit Dimension $V \geq 2$
- Und S das zugehörige Modell

Dann existiert eine globale Konstante κ, so dass

$$R_n(h, S) \geq \kappa \left[\left(\frac{V}{nh} \right) \wedge \sqrt{\frac{V}{n}} \right]$$

falls $n \geq V$

Lemma 7

- Seien $h \in [0,1]$, μ ein W-Maß auf $\mathcal{X}$ und $\mathcal{T}$ eine Menge von Klassifikatoren auf $\mathcal{X}$
- Für alle $t \in \mathcal{T}$ definieren wir P_t als W-Verteilung auf $\mathcal{X} \times \{0,1\}$ so dass unter $P_t, \mathcal{X}$ die Randverteilung μ hat und $P_t(Y|X=x) = \eta_t(x)^y(1-\eta_t(x))^{1-y}$ (Bernoulli)
- Für Partition $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2$ sei $\eta_t(x) = (1+(2t(x)-1)h)/2 \, \forall x \in \mathcal{X}_1$ und $\eta_t(x) = t(x) = 0 \, \forall x \in \mathcal{X}_2$

Für $t, s \in \mathcal{T}$ gilt dann für den Hellinger-Abstand zwischen P_t und P_s

$$\mathcal{H}^2(P_t, P_s) = (1 - \sqrt{1-h^2}) \|t - s\|_{\mathbb{L}_1(\mu)}$$

Und falls $h < 1$, beträgt die Kullback-Leibler-Information

$$\mathcal{K}(P_t, P_s) = h \log\left(\frac{1+h}{1-h}\right) \|t - s\|_{\mathbb{L}_1(\mu)}$$

Ungleichung von Assouad

- Sei $\mathcal{P}$ eine Familie von Wahrscheinlichkeitsverteilungen und
- Gebe es eine Familie $\mathcal{C} \subseteq \mathcal{P}$ mit $\#\mathcal{C} = 2^r$
- Wir schreiben $\mathcal{C} = \{P_z\}_{z \in \{0,1\}^r}$
- Weiter erfülle für $\beta > 0$ und für alle $x, y \in \{0,1\}^r$ mit $\|x - y\|_{\mathbb{L}_1(\mu)}$

$$\mathcal{H}^2(P_x, P_y) \leq \beta$$

Dann gilt für alle $n \geq 1$

$$\inf_{x \in \{0,1\}^r} \sup_{y \in \{0,1\}^r} \mathbb{E}[\ell(s_y^*, s_x^*)] \geq \frac{r}{2}(1 - \sqrt{2n(1-\beta)})$$

Beweis Theorem 4

- Haben $\mathcal{A}$ VC-Klasse, zerschmettert also eine Menge

$$\{x_1, x_2, \ldots, x_V\} \subseteq \mathcal{X}$$

- Definieren eine Verteilung μ mit Support auf $\{x_1, \ldots, x_V\}$ durch

$$\mu(x_i) = p \quad \forall 1 \leq i \leq V - 1$$

$$\mu(x_V) = 1 - p(V - 1)$$

- Mit $p \geq 0$ und $p(V - 1) \leq 1$

Beweis Theorem 4

- Haben Verteilung μ auf $\mathcal{X}$ und definieren weiter
- Für alle $b \in \{0,1\}^{V-1}$
- Eine gemeinsame Verteilung P_b auf $\mathcal{X} \times \{0,1\}$ so dass
- X die Randverteilung μ und Y gegeben $X = x$ Bernoulli-verteilt zum Parameter

$$\eta(x_i) = \frac{1}{2}(1 + (2b_i - 1)h) \quad \forall 1 \leq i \leq V - 1$$

$$\eta(x_V) = 0$$

Beweis Theorem 4

- Benötigen hier, dass $\mathcal{A}$ die Menge $\{x_1, x_2, \ldots, x_V\}$ zerschmettert und damit

$$P_b \in \mathcal{P}(h, S) \quad \forall b \in \{0,1\}^{V-1}$$

- Also ist

$$\{P_b, b \in \{0,1\}^{V-1}\}\} \subseteq \mathcal{P}(h, S)$$

- Das benötigen wir in Assouad's Lemma

Beweis Theorem 4

- Gegeben einen Schätzer $\hat{s}$ definieren wir $\hat{b}$ so dass

$$\min_{b' \in \{0,1\}^{V-1}} \|s^*_{b'} - \hat{s}\|_1 = \|s^*_{\hat{b}} - \hat{s}\|_1$$

- Dadurch und weitere Überlegungen erhalten wir eine Abschätzung

$$R_n(h, S) \geq \frac{h}{2} \inf_{\hat{b} \in \{0,1\}^{V-1}} \sup_{b \in \{0,1\}^{V-1}} \mathbb{E}[\|s^*_b - s^*_{\hat{b}}\|_1]$$

Beweis Theorem 4

- Wir verwenden Lemma 7 und erhalten für alle $b, b' \in \{0,1\}^{V-1}$

$$\mathcal{H}^2(P_b, P_{b'}) = p(1-\sqrt{1-h^2})(\sum_{i=1}^{V-1} \mathbb{1}_{b_i \neq b'_i})$$

- Das angewandt auf Assouad's Lemma bringt uns zu

$$R_n(h,S) \geq \frac{V-1}{54nh}$$

- Falls $h \geq \sqrt{(V-1)/n}$ und für den Fall $h \leq \sqrt{(V-1)/n}$ nehmen wir

$$\tilde{h} = \sqrt{(V-1)/n}$$

- Da auch $\{P_b, b \in \{0,1\}^{V-1}\}\} \subseteq \mathcal{P}(\tilde{h}, S)$ erhalten wir

$$R_n(h,S) \geq \frac{V-1}{54n\tilde{h}}$$

Beweis Theorem 4

- Haben $R_n(h, S) \geq \frac{V-1}{54nh}$, falls $h \geq \sqrt{(V-1)/n}$
- Und $R_n(h, S) \geq \frac{V-1}{54n\tilde{h}}$, falls $h \leq \sqrt{(V-1)/n}$
- Was uns zu der gewünschten Abschätzung führt

$$R_n(h, S) \geq \kappa \left[\left(\frac{V}{nh} \right) \wedge \sqrt{\frac{V}{n}} \right]$$

kombinatorische Eigenschaft von Klassen $\mathcal{A}$

Definition

Seien $D, N \in \mathbb{N}$ und $\mathcal{A} \subseteq \mathcal{P}^{\mathcal{X}}$. Wir sagen $\mathcal{A}$ besitzt die Eigenschaft $(A_{N,D})$, falls
$\exists$ Punkte $x_1, x_2, \ldots, x_N$ in $\mathcal{X}$ so dass die
Spur von $x = \{x_1, x_2, \ldots, x_N\}$ auf $\mathcal{A}$ $Tr(x) = \{A \cap x : A \in \mathcal{A}\}$
alle Teilmengen von x mit Kardinalität D enthält

- per Definition erfüllt jede VC-Klasse $\mathcal{A}$ mit Dimension V die Bedingung $(A_{V,D})$ für alle $1 \leq D \leq V$

Theorem 5

- Sei $D \geq 1$
- $\mathcal{A}$ erfülle $(A_{N,D})$ für alle $N \in \mathbb{N}$ mit $N \geq 4D$
- $h \in [0, 1)$

Dann gibt es eine globale Konstante c, so dass

$$R_n(h, S) \geq c(1-h)\frac{D}{nh}\left[1 + \log\left(\frac{nh^2}{D}\right)\right]$$

falls $h \geq \sqrt{\frac{D}{n}}$

Lemma 8

- Sei $N \geq 1$ und
- $(P_i)_{0 \leq i \leq N}$ eine Familie von Verteilungen, wobei
- $(A_i)_{0 \leq i \leq N}$ eine Familie von disjunkten Ereignissen
- Wir definieren $a = \min_{\leq i \leq N} P_i(A_i)$ und setzen
- $\bar{\mathcal{K}} = N^{-1} \sum_{i=1}^{N} \mathcal{K}(P_i, P_0)$

Dann gilt:

$$a \leq 0.71 \vee \left(\frac{\bar{\mathcal{K}}}{\ln(1+N)} \right)$$

Beweis Theorem 5

- Wir betrachten die Menge $\{x_1, x_2, \ldots, x_N\}$ die wir durch $(A_{N,D}$ erhalten
- Und gehen ähnlich vor wie im Beweis von Theorem 4
- Wählen aber statt dem Hyperwürfel

$$b \in \{0,1\}_D^N = \left\{ b \in \{0,1\}^N | \sum_{i=1}^N b_i = D \right\}$$

- Und P_b analog als gem Vert auf $\mathcal{X} \times \{0,1\}$ mit Randverteilung μ auf $\mathcal{X}$ und Bernoulli-Verteilung zum Parameter $\eta_b(x_i)$ von Y gegeben $X = x_i$ mit

$$\eta_b(x_i) = \frac{1}{2}(1 + (2b_i - 1)h) \quad \forall 1 \leq i \leq N$$

Beweis Theorem 5

- Mit der Eigenschaft $(A_{N,D})$ erhalten wir, dass

$$P_b \in \mathcal{P}(h, S) \quad \forall b \in \{0,1\}_D^N$$

- Weiter stellen wir fest, dass wir eine Menge $\mathcal{C} \subseteq \{0,1\}_D^N$ mit

$$R_n(h, S) \geq \frac{h}{2} \inf_{\hat{b} \in \mathcal{C}} \sup_{b \in \mathcal{C}} \mathbb{E}[\|s_b^* - s_{\hat{b}}^*\|_1]$$

- auch so wählen können, dass

$$R_n(h, S) \geq \frac{hD}{4N} \inf_{\hat{b} \in \mathcal{C}} \left(1 - \min_{b \in \mathcal{C}} \mathbb{P}_b(b = \hat{b})\right)$$

Beweis Theorem 5

- Wir verwenden das Lemma 8 und können zeigen, dass

$$\min_{b \in \mathcal{C}} \mathbb{P}_b(\hat{b} = b) \leq 0.71 \vee \frac{\bar{\mathcal{K}}}{\log(\#\mathcal{C})}$$

Wobei

$$\bar{\mathcal{K}} = \frac{n}{\#\mathcal{C} - 1} \sum_{b \in \mathcal{C}, b \neq b_0} \mathcal{K}(P_b, P_{b_0})$$

Für einen Punkt $b_0 \in \mathcal{C}$ und einen beliebigen Schätzer $\hat{b}$

Beweis Theorem 5

- Hatten bereits $R_n(h,S) \geq \frac{hD}{4N} \inf_{\hat{b}\in\mathcal{C}} \left(1 - \min_{b\in\mathcal{C}} \mathbb{P}_b(b=\hat{b})\right)$
- Mithilfe von Lemma 7 erhalten wir

$$\bar{\mathcal{K}} \leq \frac{4Dh^2 n}{(1-h)N}$$

- Und mit $\min_{b\in\mathcal{C}} \mathbb{P}_b(\hat{b}=b) \leq 0.71 \vee \frac{\bar{\mathcal{K}}}{\log(\#\mathcal{C})}$
- Können wir nun weiter abschätzen und erhalten nach ellenlangen Rechnungen die gewünschte Abschätzung

$$R_n(h,S) \geq c(1-h)\frac{D}{nh}\left[1 + \log\left(\frac{nh^2}{D}\right)\right]$$

Theorem 6

- Sei μ ein Wahrscheinlichkeitsmaß auf $\mathcal{X}$ und
- S ein Modell auf $\mathcal{X}$ so dass für positive Konstanten K_1, K_2, ϵ_0, r gilt

$$K_2\epsilon^{-r} \leq H_1(\epsilon, S, \mu) \leq K_1\epsilon^{-r}$$

für alle $0 < \epsilon \leq \epsilon_0$

Dann gibt es positive Konstante K in Abhängigkeit von K_1, K_2, ϵ_0, r, so dass

$$R_n(h, S, \mu) \geq K(1-h)^{1/(1+r)}[(h^{-(1-r)/(1+r)}n^{-1/(1+r)}) \wedge n^{-1/2}]$$

falls $n \geq 2$

Beweis Theorem 6

- Für jeden Klassifikator t inS setzen wir wieder analog zum Vorherigen
- $\eta_t(x) = (1 + (2t(x) - 1)h)/2$
- Und definieren P_t als die gemeinsame Verteilung auf $\mathcal{X} \times \{0, 1\}$
- Mit der Randverteilung μ von X
- Und $P_t(Y|X = x) = \eta_t(x)^y(1 - \eta_t(x)^{1-y})$ für $y \in \{0, 1\}$

Wie in den vorherigen Beweisen erhalten wir für eine Teilmenge $\mathcal{C} \subseteq \mathcal{S}$

$$R_n(h, S) \geq \frac{1}{2} \inf_{\hat{s} \in \mathcal{C}} \sup_{s \in \mathcal{C}} \mathbb{E}[\|s - \hat{s}\|_1]$$

Beweis Theorem 6

- Wir überlegen uns auf Grundlage von Yang und Barron ein Argument mit ϵ- Netzen bezüglich $\mathbb{L}_1(\mu)$
- Dazu sei $\mathcal{C}'$ ein ϵ-Netz von S
- Und für $C > 1$ sei $\mathcal{C}''$ ein $C\epsilon$-Netz von S

Jeder Punkt von $\mathcal{C}'$ muss also in einem $C\epsilon$-Ball um einen Punkt aus $\mathcal{C}''$ liegen
Demnach gilt für alle $t, t' \in \mathcal{C}$, wobei $\mathcal{C}$ ein Schnitt von $\mathcal{C}'$ mit einem solchen Ball von maximaler Kardinalität bezeichne:

$$\epsilon \leq \|t - t'\|_1 \leq 2C\epsilon$$

Und außerdem

$$\log(\#\mathcal{C}) \geq H_1(\epsilon, S\mu) - H_1(C\epsilon, S, \mu)$$

Beweis Theorem 6

- Mithilfe von Lemma 8 erhalten wir im Fall, dass $\bar{\mathcal{K}} \leq 0.71 \log(\#\mathcal{C})$

$$R_n(h, S) \geq (h\epsilon/2)(1 - 0.71)$$

- Und mit Lemma 7 für eine geschickte Wahl von C und C_1 in Abhängigkeit von K_1, K_2, r

$$\bar{\mathcal{K}} \leq \left(\frac{h^2}{1-h}\right)\epsilon$$

- Und für eine geschickte Wahl von C und C_1 in Abhängigkeit von K_1, K_2, r

$$\frac{\bar{\mathcal{K}}}{\log(\#\mathcal{C})} \leq \frac{8n}{C_1}\left(\frac{h^2}{1-h}\right)\epsilon^{1+r}$$

Beweis theorem 6

- Haben $R_n(h, S) \geq (h\epsilon/2)(1 - 0.71)$, falls $\bar{\mathcal{K}} \leq 0.71 \log(\#\mathcal{C})$
- Und $\frac{\bar{\mathcal{K}}}{\log(\#\mathcal{C})} \leq \frac{8n}{C_1} \left(\frac{h^2}{1-h} \right) \epsilon^{1+r}$

Woraus wir folgern können, dass

$$R_n(h, S) \geq (h\epsilon/2)(1 - 0.71)$$

Wann immer

$$\frac{8n}{C_1} \left(\frac{h^2}{1-h} \right) \epsilon^{1+r} \leq 0.71$$

Beweis Theorem 6

- haben $R_n(h, S) \geq (h\epsilon/2)(1 - 0.71)$, falls $\frac{8n}{C_1}\left(\frac{h^2}{1-h}\right)\epsilon^{1+r} \leq 0.71$
- Durch Umformen und der Annahme, dass $0.71 C_1/8 \leq \epsilon_0^{1+r}$

Erhalten wir mit der Wahl von

$$\epsilon = \left(\frac{0.71 C_1}{8}\right)^{1/(1+r)} h^{-2/(1+r)}(1-h)^{1/(1+r)} n^{-1/(1+r)}$$

dass $\epsilon \leq \epsilon_0$, falls $nh^2 \geq 1$, woraus wir folgern, dass

$$R_n(h, S, \mu) \geq K'(1-h)^{1/(1+r)}(h^{-(1-r)/(1+r)} n^{-1/(1+r)})$$

für eine Konstante K' in Abhängigkeit von K_1, K_2, r, ϵ_0

Beweis Theorem 6

- Haben falls $nh^2 \geq 1$

$$R_n(h, S, \mu) \geq K'(1-h)^{1/(1+r)}(h^{-(1-r)/(1+r)} n^{-1/(1+r)})$$

- Falls jedoch $nh^2 < 1$, ersetzen wir h durch $\tilde{h} = n^{1/2}$ und erhalten

$$R_n(h, S) \geq K'' \sqrt{1/n}$$

für eine Konstante K'' in Abhängigkeit von K_1, K_2, r, ϵ_0 und damit letztendlich zu der gewünschten Abschätzung

$$R_n(h, S, \mu) \geq K(1-h)^{1/(1+r)}[(h^{-(1-r)/(1+r)} n^{-1/(1+r)}) \wedge n^{-1/2}]$$