Dr. Stefan Richter
Statistische Analyse von maschinellen Lernalgorithmen
Wintersemester 2020/2021

10. Übungsblatt - Lösungen

Aufgabe 1 (Covering Numbers von Baumstümpfen)
Lösung:

(a) Mit dem Satz gilt:

$$\begin{aligned} N(\varepsilon, \mathcal{C}, \|\cdot\|_{2,n,X}) &\leq 13 \cdot \mathcal{V}(\mathcal{C}) \cdot (\frac{4e}{\varepsilon^2})^{\mathcal{V}(\mathcal{C})} \\ &= \Big(\frac{4e(13\mathcal{V}(\mathcal{C}))^{\frac{1}{2\mathcal{V}(\mathcal{C})}}}{\varepsilon}\Big)^{2\mathcal{V}(\mathcal{C})} \\ &\overset{\forall x>0: x^{1/2x}\leq 2}{\leq} \Big(\frac{13\cdot 2\cdot 4e}{\varepsilon}\Big)^{2\mathcal{V}(\mathcal{C})} \\ &\overset{2\mathcal{V}(\mathcal{C})\leq V}{\leq} \Big(\frac{13\cdot 2\cdot 4e}{\varepsilon}\Big)^{V}. \end{aligned}$$

(b) Zur Berechnung von $m_{\mathcal{C}}(N)$ bzw. zumindest einer unteren Schranke müssen überlegen, wie viele verschiedene Labelings von N Punkten im Raum durch Baumstumpf-Entscheidungsregeln erreicht werden können. Hierbei darf man sich die Lage der Punkte im Raum aussuchen (wegen dem max in $m_{\mathcal{C}}(N)$) !

Seien $x_1, ..., x_N \in [0,1]$ beliebig, alle paarweise verschieden, so ist $m_{\mathcal{C}}(x_1, ..., x_N) = 2N$, denn: Wähle

$$f_{1j}(x) = \mathbb{1}_{\{x<x_j\}} - \mathbb{1}_{\{x\geq x_j\}}, \quad f_{2j}(x) = -\mathbb{1}_{\{x<x_j\}} + \mathbb{1}_{\{x\geq x_j\}}.$$

Dann erzeugen alle $2N$ Entscheidungsregeln f_{1j}, f_{2j} jeweils verschiedene Labelings von $x_1, ..., x_N$ (Anschaulich: Lege senkrechte Gerade jeweils genau an einen Punkt x_j an; links davon alles Klasse 1, rechts davon alles Klasse -1 und umgekehrt).
Alle anderen Baumstumpf-Entscheidungsregeln erzeugen keine weiteren Labelings.
Gibt es Indizes $i, j \in \{1, ..., N\}$ mit $x_i = x_j$, so gilt $m_{\mathcal{C}}(x_1, ..., x_N) \leq 2N$ (klar).

$\Rightarrow m_{\mathcal{C}}(N) = 2N$.

(c) In d Dimensionen: Für $x^{(1)}, ..., x^{(N)} \in [0,1]^d$ paarweise verschieden wähle

$$f_{1jk}(x) = \mathbb{1}_{\{x<x_j^{(k)}\}} - \mathbb{1}_{\{x\geq x_j^{(k)}\}}, \quad f_{2jk}(x) = -\mathbb{1}_{\{x<x_j^{(k)}\}} + \mathbb{1}_{\{x\geq x_j^{(k)}\}}, \quad j = 1, ..., d, k = 1, ..., N.$$

Alle anderen Baumstumpf-Entscheidungsregeln erzeugen keine weiteren Labelings (nicht total trivial, aber in Bild offensichtlich!).
Gibt es Indizes $i, j \in \{1, ..., N\}$ mit $x_i = x_j$, so gibt es höchstens noch weniger verschiedene Labelings.

$\Rightarrow m_{\mathcal{C}}(x_1, ..., x_N) \leq \min\{2Nd, 2^N\}$ (es kann höchstens 2^N Labelings geben, und durch f_{1jk}, f_{2jk} gibt es höchstens $2Nd$).
Für $m_{\mathcal{C}}(x_1, ..., x_N) = \min\{2Nd, 2^N\}$ wäre also nur noch zu zeigen, dass $m_{\mathcal{C}}(x_1, ..., x_N) \geq \min\{2Nd, 2^N\}$, d.h. es mindestens eine Verteilung von Punkten gibt, für welche alle f_{1jk}, f_{2jk} verschiedene Labelings erzeugen. Das kann man sich graphisch überlegen, ist aber mathematisch exakt etwas aufwendig.

(d) aus (c) $\Rightarrow$

$$\begin{aligned} & m_{\mathcal{C}}(x_1, ..., x_N) \leq \min\{2Nd, 2^N\} \\ \Rightarrow \quad & \mathcal{V}(\mathcal{C}) \leq \inf\{N \in \mathbb{N} : \min\{2Nd, 2^N\} < 2^N\} = \inf\{N \in \mathbb{N} : 2Nd < 2^N\}. \end{aligned}$$

Es gilt mit $N = 2\log_2(2d)$:

$$2^N = (2d)^2 = 4d^2, \qquad 2Nd = 4d\log_2(2d),$$

d.h. wegen $d > \log_2(2d)$ für $d \geq 2$ (Hinweis):

$$2^N > 2Nd.$$

$\Rightarrow \mathcal{V}(\mathcal{C}) \leq N \overset{\mathcal{V}(\mathcal{C})\in\mathbb{N}}{\Rightarrow} \mathcal{V}(\mathcal{C}) \leq \lfloor 2\log_2(2d) \rfloor$.

Aufgabe 2 (Beweis der Margin Condition, Lemma 6.27(ii))
Lösung:

(a) Es gilt $\mathbb{E}\tilde{L}(Y, \delta(X)) = \mathbb{E}[\mathbb{E}[\phi(-Y\delta(X))|X = x]]$, und

$$\mathbb{E}[\phi(-Y\delta(X))|X = x]] = \phi(-\delta(x))\eta(x) + \phi(\delta(x))(1 - \eta(x)).$$

Punktweises Minimieren liefert $\delta^*(x) \in \operatorname{argmin}_{z\in\mathbb{R}} \Phi_{\eta(x)}(z)$ mit $\Phi_\eta(z) := \phi(-z)\eta + \phi(z)(1-\eta)$.
ϕ differenzierbar $\Rightarrow z \mapsto \Phi_{\eta(x)}(z)$ differenzierbar und daher

$$0 = \Phi'_{\eta(x)}(\delta^*(x)).$$

Aufgefasst als Gleichung für $\delta^*(x)$ ist somit $\delta^*(x) = (\Phi'_{\eta(x)})^{-1}(0)$, d.h. setze $g(\eta) = (\Phi'_\eta)^{-1}(0)$. Es gilt außerdem

$$0 = \Phi'_\eta(g(\eta)) = -\phi'(-g(\eta))\eta + \phi'(\eta)(1 - \eta), \qquad (*)$$

dies liefert die Gleichung.

(b) Es gilt (bedingen auf X):

$$\begin{aligned} & \mathbb{E}[(\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X)))^2|X = x] \\ \overset{\delta^*(x)=g(\eta(x))}{=} \quad & \eta(x)\big[\phi(-\delta(x)) - \phi(-g(\eta(x)))\big]^2 + (1 - \eta(x))\big[\phi(\delta(x)) - \phi(g(\eta(x)))\big]^2 \\ =: \quad & A_2(\eta(x), \delta(x)), \end{aligned}$$

mit

$$A_2(\eta, \delta) := \eta \cdot \big[\phi(-\delta) - \phi(-g(\eta))\big]^2 + (1 - \eta)\big[\phi(\delta) - \phi(g(\eta))\big]^2.$$

und analog

$$\begin{aligned} & \mathbb{E}[\tilde{L}(Y, \delta(X)) - \tilde{L}(Y, \delta^*(X))|X = x] \\ = \quad & \eta(x)\big[\phi(-\delta(x)) - \phi(-g(\eta((x)))\big] + (1 - \eta(x))\big[\phi(\delta(x)) - \phi(g(\eta(x)))\big] =: A_1(\eta(x), \delta(x)) \end{aligned}$$

mit

$$A_1(\eta, \delta) := \eta \cdot \big[\phi(-\delta) - \phi(-g(\eta))\big] + (1 - \eta)\big[\phi(\delta) - \phi(g(\eta))\big].$$

(c) Es gilt

$$\begin{aligned}\partial_\eta A_1(\eta,\delta) &= \big[\phi(-\delta)-\phi(-g(\eta))\big]-\big[\phi(\delta)-\phi(g(\eta))\big]\\ &\quad +\big[\eta\phi'(-g(\eta))-(1-\eta)\phi'(g(\eta))\big]\cdot g'(\eta)\\ &\overset{(*)}{=} \big[\phi(-\delta)-\phi(-g(\eta))\big]-\big[\phi(\delta)-\phi(g(\eta))\big]. \end{aligned} \tag{1}$$

(d) Außerdem

$$\begin{aligned} &\partial_\eta A_2(\eta,\delta)\\ =\ & \big[\phi(-\delta)-\phi(-g(\eta))\big]^2-\big[\phi(\delta)-\phi(g(\eta))\big]^2\\ &\quad +2\big\{\eta\phi'(-g(\eta))\big[\phi(-\delta)-\phi(-g(\eta))\big]-(1-\eta)\phi'(g(\eta))\big[\phi(\delta)-\phi(g(\eta))\big]\big\}g'(\eta)\\ =\ & \underbrace{\big\{\phi(-\delta)-\phi(-g(\eta))-\phi(\delta)+\phi(g(\eta))\big\}}_{=\partial_\eta A_1(\eta,\delta)}\\ &\quad \times\big\{\phi(\delta)-\phi(g(\eta))+\phi(-\delta)-\phi(-g(\eta))+\big\{\eta\phi'(-g(\eta))+(1-\eta)\phi'(g(\eta))\big\}g'(\eta)\big\}.\\ =\ & \partial_\eta A_1(\eta,\delta)\cdot\{\phi(\delta)+\phi(-\delta)+B(\eta)\} \end{aligned} \tag{2}$$

Hier haben wir beim ersten Summanden 3. Binomische Formel genutzt $a^2-b^2=(a-b)(a+b)$, beim zweiten

$$\begin{aligned} &2\eta\phi'(-g(\eta))\big[\phi(-\delta)-\phi(-g(\eta))\big]-2(1-\eta)\phi'(g(\eta))\big[\phi(\delta)-\phi(g(\eta))\big]\\ =\ & \underbrace{\eta\phi'(-g(\eta))}_{\overset{(*)}{=}(1-\eta)\phi'(g(\eta))}\big[\phi(-\delta)-\phi(-g(\eta))\big]-\underbrace{(1-\eta)\phi'(g(\eta))}_{\overset{(*)}{=}\eta\phi'(-g(\eta))}\big[\phi(\delta)-\phi(g(\eta))\big]\\ &\quad -(1-\eta)\phi'(g(\eta))\big[\phi(-\delta)-\phi(-g(\eta))\big]+\eta\phi'(-g(\eta))\big[\phi(\delta)-\phi(g(\eta))\big]\\ =\ & \big\{\eta\phi'(-g(\eta))+(1-\eta)\phi'(g(\eta))\big\}\cdot\big\{\phi(-\delta)-\phi(-g(\eta))-\phi(\delta)+\phi(g(\eta))\big\}. \end{aligned}$$

(e) Es gilt

$$\partial_\eta A_1(\eta,\delta)\begin{cases}\geq 0, & g(\eta)\geq\delta\\ \leq 0, & g(\eta)\leq\delta.\end{cases}$$

Falls $g(\eta)\geq\delta$, folgt also mit (d):

$$\begin{aligned}\partial_\eta A_2(\eta,\delta) &\leq \{\underbrace{\phi(\delta)+\phi(-\delta)}_{\leq\phi(\rho)+\phi(-\rho)}+\underbrace{B(\eta)}_{\leq C_\phi}\}\cdot\underbrace{\partial_\eta A_1(\eta,\delta)}_{\geq 0}\\ &\leq \{\phi(\rho)+\phi(-\rho)+C_\phi\}\cdot\partial_\eta A_1(\eta,\delta).\end{aligned}$$

Falls $g(\eta)\leq\delta$, gilt die umgekehrte Ungleichung wegen $\partial_\eta A_1(\eta,\delta)\leq 0$.
Die Ungleichung $\phi(\delta)+\phi(-\delta)\leq\phi(\rho)+\phi(-\rho)$ folgt aus der Konvexität von ϕ ($\Rightarrow\phi'$ monoton wachsend): Es ist

$$\frac{\phi(-\delta)-\phi(-\rho)}{\rho-\delta}\leq\phi'(-\delta)\leq\phi'(\delta)\leq\frac{\phi(\rho)-\phi(\delta)}{\rho-\delta}.$$

(f) Trivial. Es tritt jeweils in $A_1(g^{-1}(\delta),\delta)$ und $A_2(g^{-1}(\delta),\delta)$ auf:

$$\phi(-\delta)-\phi(-g(g^{-1}(\delta)))=0\qquad\text{bzw.}\qquad\phi(\delta)-\phi(g(g^{-1}(\delta)))=0.$$

(g) Integration $\Rightarrow$ Für $g(\eta) \geq \delta$:

$$\begin{aligned} A_2(\eta, \delta) &= \int_{g^{-1}(\delta)}^{\eta} \partial_\eta A_2(s, \delta) ds \overset{(e)}{\leq} \{\phi(\rho) + \phi(-\rho) + C_\phi\} \cdot \int_{g^{-1}(\delta)}^{\eta} \partial_\eta A_1(s, \delta) ds \\ &\leq \{\phi(\rho) + \phi(-\rho) + C_\phi\} A_1(\eta, \delta). \end{aligned}$$

Für $g(\eta) \leq \delta$ nutze $A_2(\eta, \delta) = \int_{g^{-1}(\delta)}^{\eta} \partial_\eta A_2(s, \delta) ds = -\int_{\eta}^{g^{-1}(\delta)} \partial_\eta A_2(s, \delta) ds$.

(h) Bestimmung von C_ϕ:

- Falls $\phi(x) = e^x$: $g(\eta) = \frac{1}{2}\log(\frac{\eta}{1-\eta}) \Rightarrow g'(\eta) = \frac{1}{2}(\frac{1}{\eta} + \frac{1}{1-\eta}) = \frac{1}{2\eta(1-\eta)}$. Außerdem

$$\phi'(g(\eta)) = \eta^{1/2}(1-\eta)^{-1/2}, \qquad \phi'(-g(\eta)) = \eta^{-1/2}(1-\eta)^{1/2}.$$

Einsetzen:

$$\begin{aligned} C_\phi &= 0 \vee \max_\eta \Big\{ \frac{2(\eta(1-\eta))^{1/2}}{2\eta(1-\eta)} - \eta^{1/2}(1-\eta)^{-1/2} - \eta^{-1/2}(1-\eta)^{1/2} \Big\} \\ &= 0 \vee \max_\eta \big\{ \eta^{-1/2}(1-\eta)^{-1/2} - \eta^{1/2}(1-\eta)^{-1/2} - \eta^{-1/2}(1-\eta)^{1/2} \big\} = 0. \end{aligned}$$

- Falls $\phi(x) = \log(1 + e^x)$: $g(\eta) = \log(\frac{\eta}{1-\eta}) \Rightarrow g'(\eta) = \frac{1}{\eta(1-\eta)}$. Außerdem

$$\phi'(x) = \frac{e^x}{1+e^x}, \qquad \phi'(g(\eta)) = \eta, \qquad \phi'(-g(\eta)) = 1 - \eta.$$

Einsetzen:

$$C_\phi = 0 \vee \max_\eta \Big\{ \frac{2\eta(1-\eta)}{\eta(1-\eta)} + \log(\eta(1-\eta)) \Big\} = \max_\eta \big\{ 2 + \log(\underbrace{\eta(1-\eta)}_{\leq 4^{-1}}) \big\} = 2 - 2\log(2).$$

Aufgabe 3 (Beweis Lemma 6.31 / Covering Numbers übertragen)
Lösung: Wir müssen in den Schritten (a)-(d) jeweils zeigen, dass ein $\tilde{\varepsilon}$-Covering der rechten Seite jeweils auch ein ε-Covering der linken Seite liefert. Wir geben jeweils an, wie die Coverings der linken Seite aus denen der rechten Seite konstruiert werden.

(a) Ist $(\tilde{f}_j)$ ein ε-Covering von $\mathcal{F}_1$, so ist $f_j(x, y) := \tilde{f}_j(x, y) - \tilde{L}(y, \delta_0(x))$ ein ε-Covering von $\mathcal{F}$, denn:
Ist $f \in \mathcal{F}$ beliebig, so kann f geschrieben werden als $f = \tilde{f} - \tilde{L}(y, \delta_0(x))$ mit $\tilde{f} \in \mathcal{F}_1$. Sei j mit $\|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \varepsilon$. Dann gilt:

$$\|f - f_j\|_{2,n} = \|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \varepsilon.$$

$\Rightarrow N(\varepsilon, \mathcal{F}, \|\cdot\|_{2,n}) \leq N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n})$.

(b) Ist $(\tilde{f}_j)$ ein ε-Covering von $\mathcal{F}_2$, so ist $f_j := \phi \circ \tilde{f}_j$ ein $\varepsilon \cdot \phi'(\rho)$-Covering von $\mathcal{F}_1$, denn:
Für $f \in \mathcal{F}_1$ gibt es $\tilde{f} \in \mathcal{F}_2$ mit $f = \phi \circ \tilde{f}$. Sei j mit $\|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \varepsilon$. Dann gilt

$$\begin{aligned} \|f - f_j\|_{2,n} &= \Big(\frac{1}{n}\sum_{i=1}^{n} \{\phi(\tilde{f}(X_i, Y_i)) - \phi(\tilde{f}_j(X_i, Y_i))\}^2\Big)^{1/2} \\ &\leq \phi'(\rho) \cdot \Big(\frac{1}{n}\sum_{i=1}^{n} \{\tilde{f}(X_i, Y_i) - \tilde{f}_j(X_i, Y_i)\}^2\Big)^{1/2} = \phi'(\rho)\|\tilde{f} - \tilde{f}_j\|_{2,n} \leq \phi'(\rho)\varepsilon. \end{aligned}$$

denn ϕ ist auf $[-\rho, \rho]$ Lipschitz-stetig mit Konstante $\phi'(\rho)$.
Umstellen $\Rightarrow N(\varepsilon, \mathcal{F}_1, \|\cdot\|_{2,n}) \leq N(\frac{\varepsilon}{\phi'(\rho)}, \mathcal{F}_2, \|\cdot\|_{2,n})$.

(c) Ist (δ_j) ein ε-Covering von $\mathcal{F}_3$ bzgl. $\|\cdot\|_{2,n,X}$, so ist $f_j(x,y) = y\delta_j(x)$ ein ε-Covering von $\mathcal{F}_2$ bzgl. $\|\cdot\|_{2,n}$, denn:
Für $f \in \mathcal{F}_2$ gibt es $\delta \in \mathcal{F}_3$ mit $f(x,y) = y\delta(x)$. Sei j mit $\|\delta - \delta_j\|_{2,n,X} \leq \varepsilon$. Dann gilt

$$\|f - f_j\|_{2,n} = \Big(\frac{1}{n}\sum_{i=1}^{n} \underbrace{\{Y_i\delta(X_i) - Y_i\delta_j(X_i)\}^2}_{=Y_i^2(\delta(X_i)-\delta_j(X_i))^2}\Big)^{1/2} \overset{Y_i\in\{-1,1\}}{\leq} \|\delta - \delta_j\|_{2,n,X} \leq \varepsilon.$$

$\Rightarrow N(\varepsilon, \mathcal{F}_2, \|\cdot\|_{2,n}) \leq N(\varepsilon, \mathcal{F}_3, \|\cdot\|_{2,n,X})$.

(d) Ist $(\tilde{f}_j)$ ein ε-Covering von $\mathcal{F}_4$, so ist $f_j := \rho \cdot f_j$ ein $\rho\varepsilon$-Covering von $\mathcal{F}_3$, denn: Analog zu (b) mit $\phi(x) = \rho \cdot x$.

Homepage der Vorlesung:
`https://ssp.math.uni-heidelberg.de/sam-ws2020/`