Computational Learning Theory

Sommersemester 2018

Mario Holldack, M. Sc.
Prof. Dr. Georg Schnitger
Hannes Seiwert, M. Sc.

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Institut für Informatik
AG Theoretische Informatik

Übungsblatt 5

Ausgabe: 14.05.2018
Abgabe: 22.05.2018

Wegen des Feiertags (Pfingstmontag) endet die Abgabefrist erst am Dienstag um 10:00 Uhr.

Zur Erinnerung: Die folgende Ungleichungskette gilt für jede endliche Konzeptklasse $\mathcal{C}$:

$$\mathrm{VC}(\mathcal{C}) \leq \mathrm{Tiefe}(\mathcal{C}) = \mathrm{Gegenbeispiel}(\mathcal{C}) \leq \mathrm{Gegenbeispiel}_{\mathrm{Halbierung}}(\mathcal{C}) \leq \log_2 |\mathcal{C}|.$$

Die erste Ungleichung gilt sogar für unendliche Konzeptklassen.

Aufgabe 5.1 *Das Pebble-Spiel und PAC-Lernen regulärer Sprachen* $(4 + 4 = 8$ Punkte)

Sei S ein $\{\wedge, \vee, \neg\}$-Schaltkreis mit genau n Eingabebits $x_1, \ldots, x_n$ sowie genau einem Ausgabebit $S(x_1, \ldots, x_n)$. Der zu S gehörige Graph G_S ist ein gerichteter, azyklischer Graph mit n Quellen und einer Senke. Der Eingrad von G_S ist höchstens 2 und die Tiefe, d. h. die Länge eines längsten Weges in G_S, sei t. Der Schaltkreis S *akzeptiert* die Eingabe $(x_1, \ldots, x_n)$ genau dann, wenn $S(x_1, \ldots, x_n) = 1$ ist.

a) In der Vorlesung wurde das Pebble-Spiel vorgestellt (siehe Folie 153), welches wir nun auf dem Graphen G_S spielen. Ziel des Spiels ist es, einen Pebble auf die Senke zu setzen.

   Zeigen Sie (durch vollständige Induktion): Wir benötigen höchstens $t+1$ Pebbles und höchstens $2^{t+1} - 1$ Züge, um das Pebble-Spiel auf G_S zu gewinnen.

b) Zeigen Sie:

   Es gibt einen DFA A_S über dem Alphabet $\{0, 1\}$ mit $\mathrm{poly}(2^t)$ Zuständen, sodass für alle $w \in \{0, 1\}^n$ gilt:

$$A_S \text{ akzeptiert } \underbrace{ww\ldots w}_{2^{t+1}-1\text{-mal}} \iff S \text{ akzeptiert } w$$

   Eine informelle, aber überzeugende Begründung für die Korrektheit Ihres DFAs reicht aus.

   Hinweis: Konstruieren Sie einen DFA, der sowohl die möglichen Spielsituationen und Spielzüge des Pebble-Spiels simuliert wie auch parallel dazu den Schaltkreis auswertet. Welche Informationen müssen Sie sich in den Zuständen bzw. in den Übergängen des DFA merken? Inwiefern hilft Ihnen die Argumentation über das Pebble-Spiel aus Teilaufgabe a) bei der Modellierung der Spielzüge? Der DFA soll genau dann akzeptieren, wenn ein Pebble auf die Senke gesetzt wird und der Schaltkreis eine Eins ausgibt.

Fazit: Wenn die RSA-Hypothese gilt, dann sind reguläre Sprachen nicht effizient PAC-lernbar, selbst wenn wir uns auf die vervielfachte Gleichverteilung beschränken.

Bitte wenden!

Aufgabe 5.2 *VC, Gegenbeispiele und der Halbierungsalgorithmus* $(3{\times}2+4 = 10$ Punkte)

a) Sei HALF-INTERVAL$_n := \big\{\{1, 2, \ldots, i\} : i \in \{1, \ldots, n\}\big\}$ die Konzeptklasse der Halbintervalle über dem Beispielraum $X := \mathbb{N}$. Bestimmen Sie die VC-Dimension und die Gegenbeispielzahl von HALF-INTERVAL$_n$.

b) Geben Sie eine eine Konzeptklasse $\mathcal{C} = \{c_1, \ldots, c_n\}$ an, sodass Gegenbeispiel$(\mathcal{C})$ und $\log_2(|\mathcal{C}|)$ möglichst weit auseinander liegen.

c) Betrachten Sie die Konzeptklasse REAL-HALF-INTERVAL $:= \big\{[0, a] : a \in [0, 1]\big\}$ der reellen Halbintervalle in $[0, 1]$ über dem Beispielraum $X := [0, 1]$. Zeigen Sie, dass es keinen Online-Algorithmus gibt, der jedes Konzept in REAL-HALF-INTERVAL nach endlich vielen Gegenbeispielen lernt.

Fazit: REAL-HALF-INTERVAL *ist nicht online-lernbar.*

d) Gegeben sei die Beispielraum $X := \{1, \ldots, 8\}$ und die folgende Konzeptklasse $\mathcal{C} := \{c_1, \ldots, c_9\}$:

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
1:	1	0	0	0	0	0	0	0	0
2:	0	1	0	0	0	0	0	0	0
3:	0	0	1	0	0	0	0	0	0
4:	0	0	0	1	0	0	0	0	0
5:	0	0	0	0	1	0	0	0	0
6:	0	0	0	0	0	1	1	1	1
7:	0	0	0	0	0	0	0	1	1
8:	0	0	0	0	0	0	1	0	1

Dabei steht in Zeile i und Spalte c_j genau dann eine 1, wenn $i \in c_j$ ist.

i) Der Halbierungsalgorithmus erhalte zunächst die Beispiele $6, 7$ und 8 (in dieser Reihenfolge). Zeigen Sie Gegenbeispiel$_{\text{Halbierung}}(\mathcal{C}) \geq 3$.

ii) Zeigen Sie, dass der Halbierungsalgorithmus nicht optimal ist, indem Sie Tiefe$(\mathcal{C})$=2 nachweisen, d. h. zeigen Sie, dass ein Schüler mit dem optimalen Schüleralgorithmus für jede Folge vorgelegter Beispiele des Lehrers nach höchstens zwei Gegenbeispielen gewinnt.

Aufgabe 5.3 *Threshold-Funktionen* $(3 + 2 + 3 + 4^* = 8$ Punkte $+ 4$ Extrapunkte)

a) Sei $n \geq k \geq 3$ und sei $f : \{0, 1\}^n \to \{0, 1\}$ die boolesche Funktion mit

$$f(x_1, \ldots, x_n) := (x_1 \wedge x_2) \vee x_3 \vee \cdots \vee x_k.$$

Konstruieren Sie eine zu f äquivalente Threshold-Funktion $h : \{0, 1\}^n \to \{-1, 1\}$ mit dem Schwellenwert $t = \frac{3}{4}$, d. h. es gelte $f(x) = 1 \iff h(x) = 1$ für alle $x \in \{0, 1\}^n$.

b) Geben Sie eine möglichst einfache boolesche Funktion an, für die es keine äquivalente Threshold-Funktion gibt.

c) Seien $n, W \in \mathbb{N}_{>0}$ und $S := \{-1, 1\}^n$. Gegeben sei eine boolesche Funktion $f : S \to \{-1, 1\}$, sodass eine Threshold-Funktion $\text{sign}(\langle w, x\rangle + t)$ mit ganzzahligen Gewichten $-W \leq w_1, \ldots, w_n \leq W$ und einem ganzzahligen Schwellenwert $-W \leq t \leq W$ sowie $f(x) \cdot (\langle w, x\rangle + t) > 0$ für alle $x \in S$ existiert. Bestimmen Sie eine möglichst gute untere Schranke für Margin$_f(S)$.

d*) Zeigen oder widerlegen Sie: Es gibt eine zu $f(x_1, x_2, x_3, x_4) := (x_1 \wedge x_2) \vee (x_3 \wedge x_4)$ äquivalente Threshold-Funktion.

Bitte wenden!

Aufgabe 5.4 *Modellierung: Online-Lernen mit Darwin* (3 + 3 = 6 Punkte)

Es ist eine bekannte Tatsache, dass Charles Darwin bei seiner Fahrt auf der H.M.S. Beagle die Galápagos-Inselgruppe $I := \{I_1, \dots, I_n\}$ im östlichen Pazifik erforschte. Eine besondere Laune der Natur bringt es mit sich, dass jede Insel *eindeutig* durch die auf ihr lebenden Tierarten identifiziert werden kann. Neben einigen auf *allen* Inseln vorkommenden Tieren $t_1, \dots, t_k$ gibt es auch Tiere $u_1, \dots, u_\ell$, die jeweils *nicht* auf jeder, aber *mindestens auf einer* der Inseln anzutreffen sind, d. h. jede Insel I_x besitzt eine Menge $U_x \subseteq \{u_1, \dots, u_\ell\}$ von Tieren, die in dieser Konstellation nur dort vorkommen, und es gilt $U_x \neq U_y$ für alle Inseln $I_x \neq I_y$.

Weniger bekannt ist, dass Darwin auf den Galápagos-Inseln strandete und sich die Zeit bis zur Rettung damit zu vertrieb, jede Insel nur anhand ihrer Fauna zu erkennen. Er verwendete dazu das Lernmodell des Online-Lernens. Ihm war dabei bekannt, welche Tiere auf welcher der Inseln vorkommen. Als Konzeptklasse verwendete er $\mathcal{C} = \{C_1, \dots, C_n\}$, wobei eine Insel I_x durch das Konzept

$$C_x := \{t_1, \dots, t_k\} \cup U_x$$

repräsentiert wurde. Glücklicherweise war der dort ansässige Gott Erdőseus bereit, als Lehrer zu fungieren, allerdings verlangte er für jedes Gegenbeispiel die Opferung eines Bleistifts[1].

a) Wie viele Bleistifte reichten aus, um das Online-Spiel mit Erdőseus zu gewinnen? Dabei ist davon auszugehen, dass Erdőseus möglichst viele Bleistifte geopfert bekommen wollte. Geben Sie eine möglichst gute Schranke (in Abhängigkeit von n, k oder ℓ) an und erläutern Sie kurz, welche Tiere Erdőseus zur Klassifikation vorlegen sollte.

b) Der Legende nach erschuf Erdőseus alle Tiere und verteilte sie so auf den Inseln $I_1, \dots, I_n$, dass die VC-Dimension $\mathrm{VC}(\mathcal{C})$ eine möglichst scharfe untere Schranke für die Bleistiftopferzahl darstellt. Wie tat er dies? Geben Sie eine Konzeptklasse $\mathcal{C}$ mit $\mathrm{VC}(\mathcal{C}) = \text{Bleistiftopfer}(\mathcal{C})$ an.

   Hinweis: Sie dürfen annehmen, dass n eine Zweierpotenz ist.

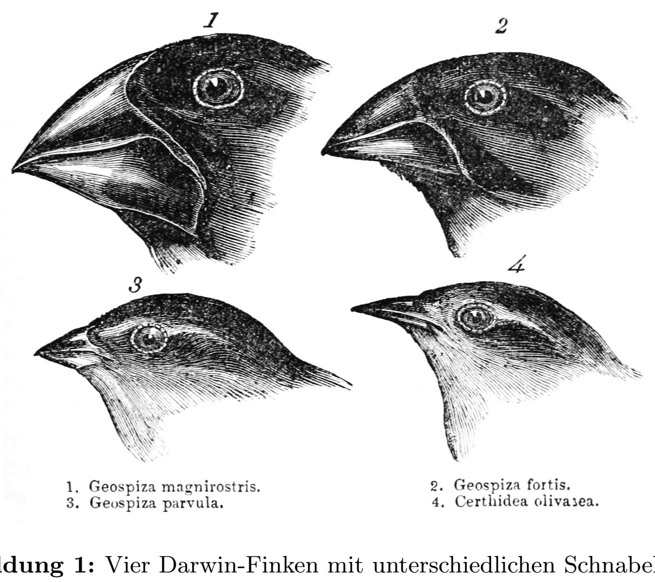

Abbildung 1: Vier Darwin-Finken mit unterschiedlichen Schnabelformen.

[1] Diese Geschichte ließ viele von Darwins Zeitgenossen an seiner Theorie zur Entstehung der Arten zweifeln.