Autor: Fabian Wleklinski
E-Mail: Fabian@Wleklinski.de
15.05.2001

Übungsblatt 1

Aufgabe 1.1:

Auf einer verstreuten Inselgruppe $I := \{I_1, I_2, \ldots, I_n\}$ im Südpazifik leben eine ganze Reihe von Tieren. Eine besondere Laune der Natur bringt es mit sich, dass abgesehen von einigen auf allen Inseln vorkommenden Tieren $t_1, t_2, \ldots t_k$ auch jede Insel I_x eine nichtleere Menge $\{u_{x,1}, u_{x,2}, \ldots\}$ von ganz spezifischen Tieren aufweist, die nur dort vorkommen.

Ein dort gestrandeter Theoretiker beschließt, sich die Zeit bis zur Rettung damit zu vertreiben, eine Insel an ihrer Fauna zu erkennen und dazu das Lernmodell des passiven Lernens zu verwenden. Ihm ist dabei bekannt, welche Tiere auf welcher der Inseln vorkommen. Als Konzept der Insel I_x setzt er also

$$C_x := \{t_1, t_2, \ldots, t_k\} \cup \{u_{x,1}, u_{x,2}, \ldots\} \tag{1.1}$$

Glücklicherweise ist eine der niederen dort ansässigen Gottheiten bereit als Lehrer zu fungieren, allerdings verlangt er für jedes Gegenbeispiel die Opferung eines Räucherstäbchens.

a. Wieviele Räucherstäbchen muß der Theoretiker bei sich haben, wenn er als Hypothesenklasse H die Konzeptklasse C (also die Menge der Inseln) verwendet? Dabei ist davon auszugehen, dass die Gottheit möglichst viele Räucherstäbchen geopfert bekommen möchte.

b. Welche Aussage macht die VC-Dimension bei diesem Problem?

c. Kann eine Hypothekenklasse angegeben werden, die die Schärfe dieser Schranke belegt?

Lösung:

Eine Anmerkung vorneweg: In einem ersten Schritt hat der Theoretiker Tiere beobachtet, diese den einzelnen Inseln zugeordnet, und ein entsprechendes Konzept für jede Insel formuliert. In dem zweiten Schritt, der in dieser Aufgabe zu lösen ist, führt der Theoretiker mit der Gottheit ein „Frage-Antwort-Spiel“, mit dem Ziel, das Lernmodell des passiven Lernens zu verwenden. Er führt dieses „Spiel“ so lange fort, bis er aus den Antworten der Gottheit zweifelsfrei schließen kann, welche Insel er beschrieben hat.

Die verschiedenen Tierarten werden durch die Variablen t_i und $u_{j,k}$ symbolisiert, siehe (1.1).

Fortan gehe ich davon aus, dass es sich bei diesen Variablen um Wahrheitswerte („*Booleans*“) handelt. Eine Variable t_i bzw. $u_{j,k}$ soll in Bezug auf das Konzept einer Insel genau dann „wahr“ sein, wenn die entsprechende Tierart auf dieser Insel vorhanden ist.

Der Theoretiker hat im Wesentlichen zwei Möglichkeiten, seine Starthypothese aufzustellen: Er kann entweder möglichst viele Literale hineinpacken, um dann von der Gottheit ein positives Gegenbeispiel zu erhalten, oder er kann möglichst wenige Literale hineinpacken, um dann ein negatives Gegenbeispiel zu erhalten. Ich gehe im Folgenden davon aus, dass er den ersten Weg wählt:

Für seine Starthypothese „verundet“ der Theoretiker die Literale aller Tierarten. (Diese Hypothese kann unmöglich richtig sein, denn dann müsste er sich ja auf allen Inseln gleichzeitig befinden.)

$$\begin{aligned} h_0 \coloneqq \quad & t_1 \wedge t_2 \wedge \ldots \wedge t_k \wedge \\ & u_{1,1} \wedge u_{1,2} \wedge \ldots \\ & u_{2,1} \wedge u_{2,2} \wedge \ldots \\ & \vdots \\ & u_{n,1} \wedge u_{n,2} \wedge \ldots \end{aligned} \tag{1.2}$$

Die Gottheit ist daran interessiert, möglichst viele Räucherstäbchen geopfert zu bekommen. Das bedeutet, dass sie möglichst oft gefragt werden möchte, den Theoretiker also so lange als möglich im Unklaren lassen möchte. Die Gottheit wird aber nicht abstreiten können, dass die Tiere $t_1,\ldots,t_k$ auf der beschriebenen Insel existieren, sowie (wenn es z.B. die Insel mit dem Index „2“ ist) die Tiere $u_{2,1},u_{2,2},\ldots$. Statt dessen wird die Gottheit Gegenbeispiele liefern, die nur leicht von der Hypothese abweichen, um dem Theoretiker nur möglichst wenige Informationen zu geben. Ein Gegenbeispiel könnte z.B. wie folgt aussehen:

$$\begin{aligned} y_0 \coloneqq \quad & t_1 \wedge t_2 \wedge \ldots \wedge t_k \wedge \\ & u_{1,1} \wedge u_{1,2} \wedge \ldots \\ & \overline{u_{2,1}} \wedge u_{2,2} \wedge \ldots \\ & \vdots \\ & u_{n,1} \wedge u_{n,2} \wedge \ldots \end{aligned} \tag{1.3}$$

In diesem Gegenbeispiel ist das Literal $u_{2,1}$ von der Gottheit negiert worden. Dadurch gibt die Gottheit dem Theoretiker zu verstehen, dass es nicht die Insel mit dem Index „2“ sein kann. Der Theoretiker wird daher alle für die Insel Nr. 2 spezifischen Tiere in seiner Hypothese negieren, so dass die Gottheit keine weiteren, negativen Gegenbeispiele mit den Literalen dieser Insel erzeugen kann. Die Gottheit hat pro Insel also nur einmal „die Möglichkeit“, ein negatives Gegenbeispiel zu präsentieren, so dass der Theoretiker nach n Gegenbeispielen (Räucherstäbchen!) zu dem Konzept gelangt ist:

$$\text{Gegenbeispiel}_C(C) = n \tag{1.4}$$

Da in der Konzeptklasse C aber insgesamt n Konzepte enthalten sind, muss der Teoretiker insgesamt n^2 Räucherstäbchen opfern.

Die VC-Dimension ist die Mächtigkeit der mächtigsten Beispielmenge, die von der Konzeptklasse zertrümmert werden kann. Daher:

$$VC(C) \coloneqq k + \sum_{i=1}^{n} m_i \tag{1.5}$$

Aufgabe 1.2:

Einem Kreis K in der Ebene $\mathbb{R}^2$ ordnen wir die Punktmenge

$$C_K \coloneqq \left\{ (x,y) \in \mathbb{R}^2 \,\middle|\, (x,y) \text{ liegt in } K \right\} \tag{1.6}$$

als Konzept zu. Sei C die Klasse aller solcher Konzepte. Bestimmen Sie exakt die VC-Dimension dieser Konzeptklasse.

Lösung:

Per Definition ist die VC-Dimension die Mächtigkeit der mächtigsten Beispielmenge S, die von C zertrümmert wird.

Es gilt daher, eine Beispielmenge zu konstruieren bzw. zu „finden“, die die Konzeptklasse aus (1.6) „zertrümmern“ kann. In diesem Fall entspricht die Beispielmenge einer Menge von Punkten, die wir uns in die selbe Ebene wie den Kreis eingezeichnet vorstellen.

Wenn unsere Beispielmenge weniger als vier Punkte enthält (inklusive der leeren Menge!), dann ist die Bedingung auf jeden Fall erfüllt, dass jede beliebige Teilmenge dieser Punkte (also null bis drei Punkte) von einem Kreis überdeckt werden kann. Dieser Zusammenhang ist ganz genereller Natur, und wird deutlich, wenn man sich einen Kreis zeichnet.

Anders ist es, wenn unsere Beispielmenge vier Punkte oder mehr enthält: Es ist nun nicht mehr möglich, einen Kreis zu zeichnen, der drei beliebige dieser vier Punkte beinhaltet. Die Begründung ist, dass dieser Kreis ja einen beliebigen dieser vier Punkte nicht beinhalten dürfte, was aber nicht in jedem Fall möglich ist.

Daher ist die VC-Dimension des Kreises: 3. (Wegen der maximalen Mächtigkeit 3).

Aufgabe 1.3:

Einem Dreieck D in der Ebene $\mathbb{R}^2$ ordnen wir die Punktmenge

$$C_D := \left\{ (x, y) \in \mathbb{R}^2 \,\middle|\, (x, y) \text{ liegt in } D \right\} \tag{1.7}$$

als Konzept zu. Sei C die Klasse aller solcher Konzepte. Bestimmen Sie exakt die VC-Dimension dieser Konzeptklasse.

Lösung:

Der Lösungsansatz ist ähnlich zu dem von Aufgabe 1.2:

Da das Dreieck eine viel „ungleichmäßigere“ Form als der Kreis besitzt, und wir es ja nach Belieben „drehen“ und „skalieren“ können, ist es uns möglich, aus bis zu sieben willkürlich gegebenen Punkten jede beliebige Teilmenge von Punkten zu erreichen.

Ich kann das leider nur anschaulich erklären, nicht rechnerisch. Eine digitale Zeichnung steht mir aber leider nicht zur Verfügung.

Mit der selben Begründung wie schon bei Aufgabe 1.2 ist daher die VC-Dimension des Dreieckes: 7. (Weil nur Teilmengen von maximal 7 Punkten erreicht werden können.)

Autor: Fabian Wleklinski 15.05.2001
E-Mail: Fabian@Wleklinski.de

Übungsblatt 2

Aufgabe 2.1:

Zur Klassifizierung von Beispielen steht uns bei den ersten beiden Aufgaben eine Runde von n Experten zur Verfügung. Diese Experten können allerdings gelegentlich irren. Folge davon ist, dass das Verfahren nicht zwangsläufig terminiert. Es macht also nur Sinn, nach der Zahl der Fehler bis zu einem beliebigen aber festen Zeitpunkt k zu fragen.

Um Fehler abzufangen, variieren wir *Weighted Majority* wie folgt:

- Zu Beginn weise jedem Experten eine *Glaubwürdigkeit* w_i von 1 zu.
- Unsere Hypothese h ist zu jedem Zeitpunkt:

$$h\left(\vec{x}\right)=1 \Leftrightarrow \sum_{E_i\left(\vec{x}\right)=1} w_i \geq \sum_{E_i\left(\vec{x}\right)=0} w_i$$

  Dabei bezeichnet $E_i\left(\vec{x}\right)$ die Klassifizierung, die Experte i bei Beispiel $\vec{x}$ vornimmt.

- Solange ein Gegenbeispiel $\vec{x}$ geliefert wird, aktualisiere die Glaubwürdigkeit der Experten wie folgt:

$$w_i := \begin{cases} w_i & \text{falls } E_i \text{ bei } \vec{x} \text{ richtig lag} \\ \dfrac{w_i}{2} & \text{sonst} \end{cases}$$

Zeigen Sie, dass nach k Gegenbeispielen für den Algorithmus auch der beste Experte mindestens $\Omega(k)-O(\log n)$ Fehler macht. Obwohl der zuverlässigste Experte nicht bekannt ist, und sich sogar mit wachsendem k ändern kann, wird unser Algorithmus also eine fast genauso gute Klassifikationsleistung zeigen.

Lösung:

Sei m die Anzahl der Fehler, die der beste Experte macht (m für „*mistake*“), und sei w die Glaubwürdigkeit dieses Experten, die ihm nach diesen Fehlern noch verblieben ist. Da die Glaubwürdigkeit eines Experten nach jedem Fehler auf die Hälfte reduziert wird, gilt

$$w=\left(\frac{1}{2}\right)^m \quad \Rightarrow \quad w=2^{-m} \quad \Rightarrow \quad \log_2 w=-m \quad \Rightarrow \quad m=-\log_2 w \tag{2.1}$$

Sei W die Summe der Glaubwürdigkeiten aller Experten(, die den Experten noch verblieben sind). Die Glaubwürdigkeit des besten Experten ist auf jeden Fall geringer als die Summe der Glaubwürdigkeiten aller Experten, d.h.

$$w \leq W \tag{2.2}$$

Immer wenn ein neues Gegenbeispiel vorliegt, wird ggf. die Glaubwürdigkeit der Experten aktualisiert. Und zwar wird die Glaubwürdigkeit jedes Experten, der eine falsche Voraussage gemacht hat, halbiert. Immer wenn der Algorithmus eine falsche Hypothese aufstellt, dann ist die Summe der Glaubwürdigkeiten aller Experten, die eine falsche Voraussage gemacht haben, mindestens die Hälfte der Summe der Glaubwürdigkeiten aller Experten; anderenfalls hätte der Algorithmus eine korrekte Hypothese aufgestellt. Dadurch wird die Summe der Glaubwürdigkeiten aller Experten (W) um mindestens ein Viertel reduziert.

k ist gemäß Aufgabenstellung die Anzahl der Gegenbeispiele, d.h. die Anzahl der falschen Hypothesen, die der Algorithmus bisher aufgestellt hat, d.h. die Anzahl seiner Fehler. Dann gilt

$$W \le n \cdot \left(\frac{3}{4}\right)^k \tag{2.3}$$

Anmerkung: Wenn der Algorithmus eine wahre Hypothese aufstellt, d.h. eine Voraussage macht, die danach nicht durch ein Gegenbeispiel wiederlegt wird, dann wird die Glaubwürdigkeit keines Experten reduziert. (Dies geht aus dem Aufgabentext hervor: „Solange ein Gegenbeispiel ...“.)

Der Rest ist Umformerei. Aus (2.1) und (2.2) folgt

$$m \ge -\log_2 W \tag{2.4}$$

, und Einsetzen von (2.3) in (2.4) ergibt

$$\begin{aligned} & m \ge -\log_2\left(n \cdot \left(\frac{3}{4}\right)^k\right) \\ & \Updownarrow \\ & m \ge -\log_2\left(\frac{3^k}{4}\right) - \log_2 n \\ & \Updownarrow \\ & m \ge k \cdot \underbrace{\log_2\left(\frac{4}{3}\right)}_{const} - \log_2 n \end{aligned} \tag{2.5}$$

Weil $k \cdot \log_2\left(\frac{4}{3}\right) = \Omega(k)$ und $\log_2 n = O(\log_2 n)$ folgt

$$m \ge \Omega(k) - O(\log_2 n) \tag{2.6}$$

, und das war zu zeigen.

Aufgabe 2.2:

In einer zweiten Variation wählen wir in jeder Iteration einen Experten zufällig und übernehmen seine Hypothese. Dabei wählen wir den Experten E_i mit Wahrscheinlichkeit

$$\frac{w_i}{\sum_{j=1}^{n} w_j}.$$

Diesmal werden wir auch bei Beispielen, die wir (also unser zufällig gewählte Experte) richtig klassifizieren, die irrenden Experten bestrafen. Klassifiziert Experte E_j das Beispiel $\vec{x}$ falsch, so multiplizieren wir seine Glaubwürdigkeit w_j mit β. Dabei gilt $0 < \beta < 1$.

Da wir den Experten zufällig wählen, macht es nur Sinn, nach der **erwarteten** Anzahl von Fehlern zu fragen.

Autor: Fabian Wleklinski 15.05.2001
E-Mail: Fabian@Wleklinski.de

Zeige, dass die asymptotische Schranke für die Zahl der Fehler des besten Experten nach k **Gegen**beispielen des Algorithmus aus Aufgabe 1 auch hier gilt. Wie verändern sich die Konstanten für verschiedene Werte von β ?

Hinweis: $\ln(1-x) \le -x$ für $x \le 1$

Lösung:

Die Ablauffolge

a. zufälliges Auswählen eines Experten

b. Befragen des Experten nach seiner Meinung und Ausgabe als Hypothese

c. Vergleichen mit dem Ereignis und Aktualisierung der Glaubwürdigkeiten

wird im Folgenden „*eine Runde*“ genannt.

Sei F_t in der Runde t die Summe der Glaubwürdigkeiten aller Experten, die in dieser Runde einen Fehler gemacht haben, dividiert durch die Summe der Glaubwürdigkeiten aller Experten insgesamt. Anders ausgedrückt: Sei F_t in der Runde t der Anteil der Glaubwürdigkeiten der irrenden Experten an der Gesamtglaubwürdigkeit:

$$F_t := \frac{\sum_{i=1}^{n} \begin{cases} w_i & \text{falls sich Experte } E_i \text{ in dieser Runde geirrt hat} \\ 0 & \text{sonst} \end{cases}}{\sum_{i=1}^{n} w_i} \tag{2.7}$$

Da sich in jeder Runde zufällig für einen der Experten entschieden wird, und dessen Voraussage bedingungslos übernommen wird, drückt F_t aus, mit welcher Wahrscheinlichkeit es in dieser Runde „zu einem Fehler hätte kommen müssen“.

Dazu ein Beispiel: Sei $n = 4$, und die Glaubwürdigkeit jedes Experten sei $w_i = 1$. In der ersten Runde entscheidet sich nur ein einziger Experte für den Wert „1“, alle drei anderen Experten für den Wert „0“. Der Algorithmus wählt aber zufällig diesen einen Experten aus, und übernimmt die Voraussage „1“ als Hypothese. Am Ende der Runde tritt der Wert „1“ ein, die Hypothese war also richtig. Da von vier Experten, die alle gleichermaßen glaubwürdig sind, drei eine falsche Voraussage gemacht haben, „hätte es in dieser Runde mit der Wahrscheinlichkeit von 75% zu einer falschen Hypothese kommen müssen“.

Vor dem Beginn der ersten Runde wird optimistischerweise davon ausgegangen, dass „wahrscheinlich“ keine Fehler passieren werden, denn es liegen bis zu diesem Zeitpunkt ja keine Informationen über Fehler aus der Vergangenheit vor.

Nach dem Ende der ersten Runde kann der in F_1 gespeicherte Wert für Voraussagen über die zweite Runde benutzt werden. Der Algorithmus geht davon aus, dass auch in der zweiten Runde wieder der in F_1 gespeicherte Anteil des Gesamtvertrauens fälschlicherweise investiert werden wird.

Der Erwartungswert für die Anzahl der Fehler nach t Runden beträgt

$$k = \sum_{j=1}^{t} F_j \tag{2.8}$$

Nach Ende der Runde t wird die Gesamtglaubwürdigkeit wie folgt aktualisiert:

$$W_t = W_{t-1}\left(1-(1-\beta)\cdot F_t\right) \tag{2.9}$$

Anmerkung: Die Formel (2.9) ist etwas komplizierter als das Äquivalent bei der ersten Aufgabe. Der Grund ist, dass bislang das β nur verwendet worden ist, um jeweils für genau einen Fehler zu bestrafen. Nun muss aufgrund des Erwartungswertes für beliebige „Fehlermengen“ bestraft werden können.

Nach t Runden beträgt die Gesamtglaubwürdigkeit:

$$W_t = n \cdot \prod_{j=1}^{t} \left(1 - (1-\beta) \cdot F_j\right) \tag{2.10}$$

Sei m die Anzahl der Fehler, die der beste Experte macht (m für „*mistake*“), und sei w die Glaubwürdigkeit dieses Experten, die ihm nach diesen Fehlern noch verblieben ist. Es ist offensichtlich, dass die Glaubwürdigkeit des besten Experten nicht größer sein kann, als die Summe der Glaubwürdigkeiten aller Experten:

$$w \overset{!}{\leq} W \tag{2.11}$$

Für die dem besten Experten verbliebene Glaubwürdigkeit nach m Fehlern gilt:

$$w = \beta^m \tag{2.12}$$

Einsetzen von (2.11) und (2.12) in (2.10) ergibt

$$\begin{aligned}
&\beta^m \leq n \cdot \prod_{j=1}^{t} \left(1 - (1-\beta) \cdot F_j\right) \\
&\Downarrow \quad \ln \\
&m \cdot \ln \beta \leq \ln n + \sum_{j=1}^{t} \ln\left(1 - (1-\beta) \cdot F_j\right) \\
&\Downarrow \quad \cdot(-1) \\
&m \cdot \ln \frac{1}{\beta} \geq -\ln n - \sum_{j=1}^{t} \ln\left(1 - (1-\beta) \cdot F_j\right) \\
&\Downarrow \quad \ln(1-x) \leq -x \text{ für } x \leq 1 \\
&m \cdot \ln \frac{1}{\beta} \geq -\ln n + (1-\beta) \cdot \sum_{j=1}^{t} F_j
\end{aligned} \tag{2.13}$$

Einsetzen von (2.8) ergibt

$$\begin{aligned}
&m \geq \frac{-\ln n + (1-\beta) \cdot k}{\ln \frac{1}{\beta}} \\
&\Downarrow \\
&m \geq \frac{(1-\beta) \cdot k}{\ln \frac{1}{\beta}} + \frac{-\ln n}{\ln \frac{1}{\beta}} \\
&\Downarrow \\
&m \geq \frac{(1-\beta) \cdot k}{\ln \frac{1}{\beta}} - \frac{\ln n}{\ln \frac{1}{\beta}}
\end{aligned} \tag{2.14}$$

Wegen

$$\underbrace{\frac{(1-\beta)}{\ln\frac{1}{\beta}}}_{const}\cdot k=\Omega(k) \tag{2.15}$$

und

$$\frac{\ln n}{\ln\frac{1}{\beta}}=\underbrace{\frac{1}{\ln\frac{1}{\beta}}}_{const}\cdot\ln n=O(\ln n) \tag{2.16}$$

gilt

$$m\geq\Omega(k)-O(\ln n)\,, \tag{2.17}$$

was zu zeigen war.

Anmerkung: $\ln(\ )$ und $\log(\ )$, sowie alle anderen Logarithmen auch, unterscheiden sich nur in einem konstanten Faktor, und verändern daher die Größenordnung nicht.

Aufgabe 2.3:

Eine totale Ordnung auf n Elementen soll gelernt werden. Dabei schließen wir Gleichheit aus. D.h. es gilt stets

$$(x,y)\in R\Leftrightarrow(y,x)\notin R\,.$$

Eine totale Ordnung repräsentieren wir durch die Menge aller geordneten Paare. Ein Beispiel ist also ein geordnetes Paar zusammen mit der Information, ob dieses Paar zur zu lernenden totalen Ordnung gehört oder nicht.

a. Bestimmen Sie die VC Dimension dieser Konzeptklasse.

b. Zeigen Sie, dass jeder Lernalgorithmus mindestens $\Omega(n\cdot\log n)$ Gegenbeispiele anfordern wird, wenn vollständige Ordnungen auf n Elementen zu lernen sind.

Lösung:

Erst einmal ein paar allgemeine Anmerkungen vorweg: Sei M die Menge der n Elemente, über denen die Ordnung definiert ist, die gelernt werden soll. Seien $\{x_1,x_2,\ldots,x_n\}$ die Elemente, dann ist jede der $n!$ Permutationen der Form $\left(x_{i_1},x_{i_2},\ldots,x_{i_n}\right)$ mit $\left(i_j\neq i_k\right)\Leftrightarrow(i\neq j)$ eine totale Ordnung.

Um eine Ordnung darzustellen/abzubilden/zu repräsentieren kann man z.B. eine Menge von Tupeln wählen, die jeweils aus zwei Elementen der Menge M bestehen:
$\left\{\left(x_{j_1},x_{k_1}\right),\left(x_{j_2},x_{k_2}\right),\ldots,\left(x_{j_m},x_{k_m}\right)\right\}$.

Wieviele dieser Tupel besitzt diese Art von Darstellung einer bestimmten Ordnung?

Es sind $m=\binom{n}{2}=\frac{n!}{2!(n-2)!}$ viele Tupel.

Jede dieser totalen Ordnungen ist ein Konzept, die Menge der $n!$ verschiedenen Ordnungen ist die Konzeptklasse C_n. v sei die noch zu bestimmende VC-Dimension der Konzeptklasse

für ein bestimmtes n. Von der Definition der VC-Dimension her ist bekannt, dass es dann (mindestens) eine „Beispielmenge“ S geben muss, die aus v Elementen (Tupeln aus $M \times M$) besteht, und die von der Konzeptklasse „*zertrümmert*“ wird.

„*Zertrümmern*“ bedeutet dabei, dass die Potenzmenge der Beispielmenge S durch die Vereinigung aller Schnittmengen von jeweils einem Konzept (einer Ordnung) und der Beispielmenge S:

$$P(S) = \{c \cap S \,|\, c \in C_n\} \tag{2.18}$$

Die Operation aus (2.18) erzeugt eine Menge, die nicht mächtiger sein kann als die größere (mächtigere) der beiden Mengen. Die zu erzeugende Menge, nämlich die Potenzmenge von S, besitzt aber $2^{|S|}$ Elemente. Von der Konzeptklasse C_n ist aber bekannt, dass sie „nur“ $n!$ Konzepte beinhaltet.

Damit ist eine obere Schranke für die VC-Dimension gefunden:

$$\begin{aligned} & 2^{|S|} \le n! \\ & \Downarrow \\ & |S| \le \log_2 n! \\ & \Downarrow \\ & VC(C_n) \le \log_2 n! \end{aligned} \tag{2.19}$$

Autor: Fabian Wleklinski 15.05.2001
E-Mail: Fabian@Wleklinski.de

Übungsblatt 3

Aufgabe 3.1:

Wir betrachten das Spiel Schiffeversenken. Gespielt wird auf einem Spielfeld mit 10 mal 10 Zellen. Ein Spieler hat vier Schiffe, der Größe 1x4, 1x3, 1x2 **und** 1x1. Er darf die Schiffe auf den Zellen beliebig waagrecht oder senkrecht (nicht diagonal) positionieren, wobei sich die Schiffe in keiner Weise berühren dürfen.

Ein Konzept sei eine mit den Regeln vereinbare Anordnung der vier Schiffe. Wir können das Konzept durch die Menge der von den Schiffen abgedeckten Zellen repräsentieren. Die Gesamtheit aller legalen Anordnungen ist eine Konzeptklasse.

Geben Sie eine möglichst gute untere Schranke für die VC-Dimension dieser Konzeptklasse an.

Lösung:

Ein Konzept ist in diesem Fall eine „gültige“ Anordnung der Schiffe. Wenn man die in der Aufgabenstellung vorgeschlagene „Kodierung“ verwendet, dann besteht ein Konzept maximal aus 10 Elementen (mehr Kästchen können die Schiffe nicht auf dem Spielplan belegen).

Es steht damit fest, dass die VC-Dimension dieser Konzeptklasse nicht größer als 10 sein kann. Warum? Wäre sie größer als 10, müsste es eine Beispielmenge mit mehr als 10 Elementen (z.B. 11 Elementen) geben, die von der Konzeptklasse zertrümmert werden kann. Das wiederum ist aber ein Widerspruch, da jedes Konzept nur aus 10 Elementen besteht, und mindestens ein Element der Potenzmenge der Beispielmenge nicht erreicht werden könnte.

Ebenso fest steht, dass die VC-Dimension dieser Konzeptklasse mindestens 4 sein muß – da jede Beispielmenge von 1, 2, 3 oder 4 Elementen zertrümmert werden kann (durch ein Schiff der entsprechenden Länge/Größe).

Das Problem ist nun darauf reduziert, für die VC-Dimension einen Wert zwischen 4 und 10 (jeweils inklusive) festzulegen ;-) Also:

$$\mathrm{VC}(C) \geq 4 . \tag{3.1}$$

Aufgabe 3.2:

In dieser Aufgabe betrachten wir die Konzeptklassen *Intervall*$_n$ und *Halbintervall*$_n$. Die Konzeptklasse *Halbintervall*$_n$ besteht dabei aus den n Konzepten

$$C_i := \{x \in \mathbb{N} \mid x \leq i\} \text{ für } 1 \leq i \leq n .$$

Die Konzeptklasse *Intervall*$_n$ besteht aus $\frac{1}{2} \cdot n \cdot (n+1)$ Konzepten

$$C_{i,j} := \{x \in \mathbb{N} \mid i \leq x \leq j\} \text{ für } 1 \leq i \leq j \leq n .$$

Bestimmen Sie die Lernkomplexitäten dieser beiden Konzeptklassen möglichst exakt (also die minimale Anzahl der Gegenbeispiele).

Lösung:

Eine Möglichkeit, die Anzahl der Gegenbeispiele zu bestimmen, führt über die *VC-Dimension*. Denn es gilt:

$$\text{Gegenbeispiel}_H(C) \geq \mathrm{VC}(C) \tag{3.2}$$

a. *Halbintervall*$_n$

Zuerst gilt es zu zeigen, dass *Halbintervall*$_n$ die VC-Dimension 1 hat. Ist dieser Schritt erledigt, kann dasselbe für die VC-Dimension 2 gezeigt werden, und so weiter...

Man wählt sich ein beliebige, einelementige Beispielmenge S, z.B. $S=\{i\}$. Es ist nun geradezu trivial, jeweils ein Konzept zu finden, dass dieses i beinhaltet (z.B. c_i), und ein weiteres, welches dies nicht tut (z.B. c_{i-1}). S ist zertrümmert, und damit bewiesen, dass die VC-Dimension von *Halbintervall*$_n$ mindestens 1 ist.

Man wählt sich nun eine zweielementige Beispielmenge S, z.B. $S=\{i_1,i_2\}$, und sei o.B.d.A. $i_1<i_2$. Es ist nun wieder sehr leicht, ein Konzept zu finden, dass keines der Elemente, oder nur das erste der Elemente aus der Beispielmenge beinhaltet (z.B. c_{i_1-1} bzw. c_{i_1}). Wenn man das Konzept c_{i_2} betrachtet, sieht man, dass es sowohl i_1 als auch i_2 beinhaltet. Es gibt also kein Konzept, dass i_1 zwar nicht beinhaltet, aber i_2 beinhaltet. S wird nicht zertrümmert.

Es gilt also:

$$\mathrm{VC}(C)\geq 1. \tag{3.3}$$

b. *Intervall*$_n$

Der Erkenntnisweg gestaltet sich ähnlich zu dem des *Halbintervall*$_n$:

Man wähle sich auch hier eine einelementige Beispielmenge S, z.B. $S=\{i\}$. Dass diese zertrümmert werden kann, ist relativ klar, denn analog zum *Halbintervall*$_n$ gibt es z.B. die Konzepte $c_{i,i}$ bzw. $c_{i-1,i-1}$, welches das Element überdeckt, bzw. dies nicht tut. Die VC-Dimension ist also zumindest 1.

Man wählt sich nun erneut eine zweielementige Beispielmenge S, z.B. $S=\{i_1,i_2\}$, und sei wieder o.B.d.A. $i_1<i_2$. Es ist nun wieder sehr leicht, ein Konzept zu finden, dass keines der Elemente, oder nur das erste der Elemente aus der Beispielmenge beinhaltet (z.B. c_{1,i_1-1} bzw. c_{1,i_1}). Es ist nun aber ebenfalls sehr leicht, ein Konzept zu finden, das nur das zweite der beiden Elemente beinhaltet (z.B. $c_{i_2,n}$). Ein Konzept, dass sowohl i_1 als auch i_2 beinhaltet, ist z.B. $c_{1,n}$. S ist zertrümmert, und damit bewiesen, dass die VC-Dimension von *Intervall*$_n$ mindestens 2 ist.

Man wählt sich nun eine dreielementige Beispielmenge S, z.B. $S=\{i_1,i_2,i_3\}$, und sei o.B.d.A. $i_1<i_2<i_3$. Es bedarf nun keiner weiteren Erklärung, dass es einfach möglich ist, Konzepte zu finden, die keines, nur eines, oder alle dieser drei Elemente beinhalten. Genauso ist es kein Problem, Konzepte zu finden, die i_1 und i_2 *oder* i_2 und i_3 beinhalten. Es ist aber nicht möglich, ein Konzept zu finden, welches i_1 und i_3 beinhaltet, aber nicht i_2. Anders ausgedrückt: Jedes Intervall, dass sowohl i_1 als auch i_3 beinhaltet, beinhaltet automatisch auch i_2. Es gilt also:

$$\mathrm{VC}(C)\geq 2. \tag{3.4}$$

Autor: Fabian Wleklinski 15.05.2001
E-Mail: Fabian@Wleklinski.de

Aufgabe 3.3:

In der Vorlesung wurde die Komplexität einer Beispielmenge S für eine *Tresholdfunktion* f wie folgt definiert:

$$K(S) := \sup\left\{ \frac{\text{Margin}(S,w,t)}{\|w\|} \middle| (w,t) \text{ ist eine Implementierung von } f \text{ auf der Menge } S \right\}$$

Bestimmen Sie **exakt** die Komplexität der booleschen Und-Funktion:

$$f(x) = 1 \quad \Leftrightarrow \quad x_1 \wedge x_2 \wedge \ldots \wedge x_n$$

Lösung:

Die boolesche Funktion kann als Treshold-Gatter betrachtet werden. Sie nimmt nur dann den Wert 1 an, wenn alle Literale der Funktion wahr sind:

$$\begin{aligned} & f(x) = 1 \quad \Leftrightarrow \quad x_1 \wedge x_2 \wedge \ldots \wedge x_n \\ & \Updownarrow \\ & f(x) = \begin{cases} 1 & \sum_{j=1}^{n} w_j x_j + t \geq 0 \\ -1 & \text{sonst} \end{cases} \end{aligned} \tag{3.5}$$

Es gibt viele Möglichkeiten, die w_j zu definieren – aber in Abhängigkeit von der gewählten Möglichkeit ergibt sich jeweils ein anderer Wert für t. Jede wählbare Möglichlichkeit von w_j und einem t nennt man eine „*Implementierung*“. Im Skript zur Vorlesung ist beispielsweise die Implementierung $w = (1,\ldots,1)$ und $t = -n + 0{,}5$ gewählt worden.

Daraus ist der folgende Zusammenhang ersichtlich, der später noch benötigt wird:

$$f(x) = 1 \quad \Rightarrow \quad \sum_{j=1}^{n} w_j x_j + t \geq 0 \tag{3.6}$$

$$f(x) = -1 \quad \Rightarrow \quad \sum_{j=1}^{n} w_j x_j + t < 0 \tag{3.7}$$

Autor: Fabian Wleklinski 15.05.2001
E-Mail: Fabian@Wleklinski.de

Gesucht ist die Komplexität der booleschen Funktion, und die ergibt sie gemäß Skript wie folgt:

$$K(S) := \sup\left\{ \frac{\text{Margin}(S,w,t)}{\|w\|} \middle| (w,t) \text{ ist eine Implementierung von } f \text{ auf der Menge } S \right\}$$

$$\Downarrow \quad \text{Margin}(S,w,t) = \min\left\{ \text{Margin}(x_i,b_i,w,t) \middle| 1 \le i \le m \right\}$$

$$K(S) := \sup\left\{ \frac{\min\left\{ \text{Margin}(x_i,b_i,w,t) \middle| 1 \le i \le m \right\}}{\|w\|} \middle| (w,t) \text{ ist eine ...} \right\}$$

$$\Downarrow \quad \text{Margin}(x_i,b_i,w,t) = b_i \cdot \left(\langle w, x_i \rangle + t \right)$$

$$K(S) := \sup\left\{ \frac{\min\left\{ b_i \cdot \left(\langle w, x_i \rangle + t \right) \middle| 1 \le i \le m \right\}}{\|w\|} \middle| (w,t) \text{ ist eine ...} \right\}$$

$$\Downarrow \quad \|w\| = \sqrt{\sum_{j=1}^{n} {w_j}^2}$$

$$K(S) := \sup\left\{ \frac{\min\left\{ b_i \cdot \left(\langle w, x_i \rangle + t \right) \middle| 1 \le i \le m \right\}}{\sqrt{\sum_{j=1}^{n} {w_j}^2}} \middle| (w,t) \text{ ist eine ...} \right\} \tag{3.8}$$

Nun geht es darum, das Supremum und das Minimum aus dem Term zu entfernen, ich entferne im Folgenden zuerst das Minimum, da es innerhalb des Supremums liegt. Was ist also $\min\left\{ b_i \cdot \left(\langle w, x_i \rangle + t \right) \middle| 1 \le i \le m \right\}$?

Nun, w und t sind im Kontext betrachtet „feste“ Größen, die der Minimum-Operator nicht beeinflussen kann. Es werden lediglich x_i und b_i variiert („laufen gelassen“), um das Minimum zu finden. Es gibt die zwei Fälle, dass x_i die boolesche Funktion erfüllt, und den gegenteiligen Fall.

a. x_i erfüllt die boolesche Funktion.

Dann gilt $b_i = 1$, und aus (3.6) folgt:

$$b_i \cdot \left(\sum_{j=1}^{n} w_j x_{i,j} + t \right) = \sum_{j=1}^{n} w_j + t \tag{3.9}$$

b. x_i erfüllt die boolesche Funktion **nicht**.

In diesem Fall ist $b_i = -1$, und aus (3.7) folgt:

$$b_i \cdot \left(\sum_{j=1}^{n} w_j x_{i,j} + t \right) = -\left(\sum_{j=1}^{n} w_j x_{i,j} + t \right) \tag{3.10}$$

Da $x_{i,j} \in \{0,1\}$, gilt:

$$\begin{aligned} &\sum_{j=1}^{n} w_j x_{i,j} \le \sum_{j=1}^{n} w_j \\ &\Downarrow \\ &\sum_{j=1}^{n} w_j x_{i,j} + t \le \sum_{j=1}^{n} w_j + t \\ &\Downarrow \\ &-\left(\sum_{j=1}^{n} w_j x_{i,j} + t\right) \ge -\left(\sum_{j=1}^{n} w_j + t\right) \end{aligned} \tag{3.11}$$

.

Das $\min\left\{b_i \cdot \left(\langle w, x_i \rangle + t\right) \middle| 1 \le i \le m\right\}$ tritt also für den Fall ein, dass x_i die boolesche Funktion erfüllt, es nimmt dann den Wert $\sum_{j=1}^{n} w_j + t$ an.

(3.8) vereinfacht sich dann zu:

$$K(S) := \sup\left\{ \frac{\sum_{j=1}^{n} w_j + t}{\sqrt{\sum_{j=1}^{n} w_j^{\,2}}} \middle| (w,t) \text{ ist eine Implementierung von } f \text{ auf der Menge } S \right\} \tag{3.12}$$

Man kann nun sehen, dass die Komplexität $K(S)$ aller Implementierungen nur vom Gewichtsvektor w und vom Schwellwert t abhängig ist. Was zu zeigen war.

Für den im Skript angeführten Fall ($w = (1,\ldots,1)$ und $t = -n + 0{,}5$) ergibt sich z.B.:

$$K(S) := \sup\left\{ \frac{\sum_{j=1}^{n} 1 + \left(-n + \frac{1}{2}\right)}{\sqrt{\sum_{j=1}^{n} 1^2}} \right\} = \sup\left\{ \frac{n - n + \frac{1}{2}}{\sqrt{n}} \right\} = \sup\left\{ \frac{1}{2\sqrt{n}} \right\} = \frac{1}{2\sqrt{n}} = \Omega\left(\frac{1}{\sqrt{n}}\right) \tag{3.13}$$

Aufgabe 3.4:

Zeige, dass der Perzeptron-Algorithmus höchstens $\text{poly}(n \cdot W)$ Gegenbeispiele benötigt, falls die zu lernende Thresholdfunktion eine Implementierung mit n ganzzahligen Gewichten $w_1,\ldots,w_n$ und einem ganzzahligen Schwellwert t besitzt, wobei $|w_i| \le W$ für jedes Gewicht.

Lösung:

Aus Satz 2.14 des Skriptes ist bekannt, dass der Perzeptron-Algorithmus nach höchstens $\left(\frac{2 \cdot R}{K(S)}\right)^2$ vielen Gegenbeispielen erfolgreich lernt. Dabei ist R die Norm, welche alle Beispiele aus S maximal besitzen dürfen.

In der Aufgabenstellung ist leider keine Einschränkung bezüglich des Wertebereiches der Beispiele gegeben, dafür ist aber eine Einschränkung für die Gewichte vorgegeben, denn jedes Gewicht muss kleiner oder gleich W sein.