3 Optimale Fehlerschranken

Offensichtlich bedeutet erfolgreiches Lernen im Mistake-Bound Modell möglichst wenig Vorhersagefehler zu machen. Wir betrachten nun die minimale Anzahl der Vorhersagefehler, die jeder Lernalgorithmus für eine gegebene Konzeptklasse C machen muß. Dabei kann C eine beliebige endliche Konzeptklasse sein. Wir lassen die einem Lernalgorithmus benötigten Ressourcen wie Zeit und Platz außer Acht.

Definition. *Sei C eine endliche Konzeptklasse über einem Universum X. Die* optimale Fehlerschranke *für C ist*

$$opt(C) = \min_{\textit{Lernalg. } A} M_A(C),$$

wobei $M_A(C)$ die maximale Anzahle der Vorhersagefehler bezeichnet, die ein Lernalgorithmus A für alle Konzepte in C und Beispielfolgen $x_1, x_2, \dots \in X$ macht.

Ein einfacher Lernalgorithmus kann für jeden Vorhersagefehler mindestens ein Konzept aus C eliminieren. Somit ist $opt(C)$ für endliche Konzeptklassen C wohldefiniert (und höchsten $|C|$), auch wenn $M_A(C)$ nicht für alle Lernalgorithmen A endlich ist.

3.1 Der Halving Algorithmus

Sei C eine endliche Konzeptklasse über einem Universum X. Mit Hilfe des sogenannten *Halving Algorithmus* können wir die bereits erwähnte triviale obere Schranke $opt(C) \leq |C|$ deutlich verbessern. Wir benutzten die Notationen

$$\xi_0(V, x) = \{c \in C \mid c(x) = 0\} \quad \text{und} \quad \xi_1(V, x) = \{c \in C \mid c(x) = 1\}$$

für beliebige Teilmengen $V \subseteq C$ und Beispiele $x \in X$. Damit können wir den Halving Algorithmus wie folgt beschreiben.

Der *Halving* Algorithmus

- Initialisiere $V_0 = C$
- Im i-ten Versuch mit Beispiel x_i treffe die Vorhersage für x_i gemäß dem Mehrheitvotum aller Konzepte in V_{i-1}, und entferne aus V_{i-1} alle Konzepte, die nicht mit der Klassifikation von x_i durch das Zielkonzept c übereinstimmen. D.h.
  - treffe Vorhersage $\begin{cases} 1 & \text{falls } |\xi_0(V_{i-1}, x_i)| \leq |\xi_1(V_{i-1}, x_i)| \\ 0 & \text{sonst} \end{cases}$
  - $V_i = \xi_{c(x_i)}(V_{i-1}, x_i)$.

Theorem 3.1. *Für jede nichtleere endliche Konzeptklasse C gilt $M_{Halving}(C) \leq \lceil \log |C| \rceil$.*

Beweis. Bei jedem Vorhersagefehler wird wenigstens die Hälfte der Konzepte in V_{i-1} eliminiert. □

Das folgende Beispiel zeigt, daß der Halving Algorithmus deutlich weniger als $\lceil \log |C| \rceil$ Vorhersagefehler machen kann.

Beispiel 3.2. Sei C die Konzeptklasse über dem Universum $X = \{0, 1\}^n$, die aus allen einelementigen Konzepten $c_x = \{x\}$ mit $x \in \{0, 1\}^n$ besteht. Für diese Konzeptklasse C macht der Halving Algorithmus immer die Vorhersage 0, und somit $M_{Halving}(C) = 1$. Es gilt aber $\log |C| = \log n$.

Aufgrund der Definition der optimalen Fehlerschranke ist $opt(C) \leq M_{Halving}(C)$ und somit folgt aus Theorem 3.1 unmittelbar die folgende obere Schranke für $opt(C)$.

Korollar 3.3. *Für jede nichtleere endliche Konzeptklasse C gilt $opt(C) \leq \lceil \log |C| \rceil$.*

Wir zeigen nun mit Hilfe des folgenden Beispiels, daß der Halving Algorithmus nicht optimal ist.

Beispiel 3.4. Sei $C = \{c_1, \ldots, c_9\}$ die Repräsentationsklasse über dem Universum $X = \{a_1, \ldots, a_8\}$, die durch die folgende Tabelle gegeben ist.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
a_1	1	0	0	0	0	0	0	0	0
a_2	0	1	0	0	0	0	0	0	0
a_3	0	0	1	0	0	0	0	0	0
a_4	0	0	0	1	0	0	0	0	0
a_5	0	0	0	0	1	0	0	0	0
a_6	0	0	0	0	0	1	1	1	1
a_7	0	0	0	0	0	0	0	1	1
a_8	0	0	0	0	0	0	1	0	1

Es gibt einen Algorithmus, der C mit höchstens zwei Vorhersage erlernt. Der Halving Algorithmus macht aber mindestens drei Vorhersagefehler.

3.2 Ein randomisierter Halving Algorithmus

...

3.3 Ein optimaler Algorithmus

Die Definition von $opt(C)$ erlaubt, daß der Algorithmus, der die optimalen Fehlerschranke realisiert, von der in Frage stehenden Konzeptklasse C abhängen darf. Wir zeigen nun, daß es tatsächlich einen Algorithmus gibt, der für jede Konzeptklasse C die optimalen Fehlerschranke $opt(C)$ realisiert. Hierzu benützen wir den folgenden Begriff eines Fehlerbaums für eine Konzeptklasse C.

Definition. *Sei C eine endliche Konzeptklasse über einem Universum X. Ein* Fehlerbaum *für C ist ein Binärbaum mit Knoten $V \subseteq C$, $V \neq \emptyset$, so daß gilt:*

1. *Die Wurzel ist C.*
2. *Alle internen Knoten sind zusätzlich mit einem Beispiel $x \in X$ markiert.*

3. *Das linke Kind eines Knotens V mit Markierung x, falls vorhanden, ist $\xi_0(V, x)$, und das rechte Kind, falls vorhanden, ist $\xi_1(V, x)$.*

Ein vollständiger k-Fehlerbaum *ist ein vollständiger Fehlerbaum der Höhe k. $K(C)$ bezeichnet die maximale Höhe eines vollständiger k-Fehlerbaums für C, falls $C \neq \emptyset$. Wir setzen $K(\emptyset) = -1$.*

Beispiel. $C = \{c_1, \ldots, c_5\}$ über $X = \{0,1\}^5$ mit $c_i(x_1, \ldots, x_5) = x_i$. Bild...

Das folgende Lemma besagt, daß jeder Algorithmus mindestens $K(C)$ Vorhersagefehler macht.

Lemma 3.5. *Für jede endliche Konzeptklasse C gilt $opt(C) \geq K(C)$.*

Beweis. Sei $K(C) = k$. Wir müssen zeigen, daß daraus $M_A(C) \geq k$ für alle Lernalgorithmen A folgt. Für einen beliebigen Lernalgorithmus A Können wir eine Beispielfolge $x_1, \ldots, x_k$ und ein Zielkonzept c so gewählt, daß

- $x_1, \ldots, x_k$ die Markierungen eines Pfades im vollständigen k-Fehlerbaum sind, bei dem A stets einen Fehler macht, und
- c ist ein Konzept im Blatt des dazugehörigen Pfades ist.

Dann macht A mit Zielkonzept c genau k Fehler, also ist $M_A(C) \geq k$. □

Der optimale Algorithmus *Opt*

- Initialisiere $V_0 = C$
- Im i-ten Versuch:
  - treffe Vorhersage $\begin{cases} 1 & \text{falls } K(\xi_0(V_{i-1}, x_i)) \leq K(\xi_1(V_{i-1}, x_i)) \\ 0 & \text{sonst} \end{cases}$
  - $V_i = \xi_{c(x_i)}(V_{i-1}, x_i)$.

Lemma 3.6. *Sei $x_1, \ldots, x_t$ eine beliebige Beispielfolge. Dann gilt für alle $k \geq 0$ und $i \in \{1, \ldots, t\}$, falls $K(V_{i-1}) = k$ dann macht Opt höchstens k Vorhersagefehler in den Versuchen $i, \ldots, t$.*

Beweis. ...

□

Theorem 3.7. *Für jede endliche Konzeptklasse C gilt $M_{OPT}(C) = opt(C)$*

Beweis. Es gilt $opt(C) \leq M_{OPT}(C)$, und aus Lemma 3.5 und 3.6 folgt

$$M_{OPT}(C) \leq K(C) \leq opt(C)$$

□

Bemerkung 3.8. Insbesondere gilt also auch $K(C) = opt(C)$.

3.4 VC Dimension

...wir betrachten nun auch unendliche Konzeptklassen.

Definition. *Sei C eine Konzeptklasse über einem Universum X. Dann wird eine Menge $S \subseteq X$ von C* erschlagen, *falls*

$$\{c \cap S \mid c \in C\} = \mathcal{P}(S).$$

Die Vapnik-Chervonenkis (*VC*) Dimension *von C ist*

$$VC(C) = \max\{|S| \mid S \subseteq X \text{ und } S \text{ wird von } C \text{ erschlagen}\},$$

falls das Maximum existiert, und $VC(C) = \infty$ sonst.

Bemerkung 3.9. Eine Konzeptklasse C erschlägt also eine Menge $S \subseteq X$, falls

$$\forall U \subseteq S\ \exists c \in C : U = c \cap S.$$

Eine *untere* Schranke $VC(C) \geq k$ ist äquivalent mit der *Existenz* einer Menge $S \subseteq X$ der Größe k die von C erschlagen wird, und eine *obere* Schranke $VC(C) < k$ erfordert den Nachweis, daß *keine* Menge $S \subseteq X$ der Größe k von C erschlagen wird.

Bemerkung 3.10. Die folgenden Eigenschaften der VC-Dimension folgen unmittelbar aus der Definition.

1. $C_1 \subseteq C_2 \implies VC(C_1) \leq VC(C_2)$
2. $VC(C) \leq \lfloor \log |C| \rfloor$ für endliche Konzeptklassen C
3. $VC(C) = VC(co\text{-}C)$

Dabei bezeichnet $co\text{-}C$ die Menge der Komplemente der Konzepte einer Konzeptklasse C über einem Universum X, $co\text{-}C = \{X - c \mid c \in C\}$.

Beispiel 3.11 (halboffene Intervalle). Sei

$$C = \{(-\infty, x] \mid x \in \mathbb{R}\} \cup \{[x, \infty) \mid x \in \mathbb{R}\}$$

die Konzeptklasse der halboffenen Intervalle über dem Universum $X = \mathbb{R}$. C erschlägt die Menge $S = \{1, 2\}$ (tatsächlich jede Menge $\{x_1, x_2\}$ mit $x_1 \neq x_2$), also ist $VC(C) \geq 2$. Für eine obere Schranke von $VC(C)$ beobachten wir, daß keine Menge $S = \{x_1, x_2, x_3\}$ mit $x_1 < x_2 < x_3$ durch halboffene Intervalle erschlagen werden kann, d.h. für $U = \{x_1, x_3\} \subseteq S$ gibt es kein $x \in \mathbb{R}$ mit $U = (-\infty, x] \cap S$ oder $U = [x, \infty) \cap S$. Also ist $VC(C) = 2$.

Beispiel 3.12 (Halbebenen in $\mathbb{R}^2$). Sei C die Konzeptklasse aller Halbebenen

$$c = \{(x, y) \in \mathbb{R}^2 \mid ax + b \leq y\}$$

in der Euklidischen Ebene $\mathbb{R}^2$, Drei Punkte $P_1, P_2, P_3 \in \mathbb{R}^2$, die nicht auf einer Linie liegen, werden (immer) erschlagen. Also gilt $VC(C) \geq 3$. Vier Punkte P_1, P_2, P_3, P_4 werden aber nie erschlagen:

1. Falls P_1, P_2, P_3, P_4 hintereinander auf einer Linie liegen, dann wählen wir $U = \{P_1, P_3\}$.
2. Falls P_1, P_2, P_3, P_4 die Eckpunkte eines konvexen Polygons bilden, dann können wir $U = \{P_1, P_3\}$ für zwei gegenüberliegenden Punkte wählen.
3. Falls es einen “inneren Punkt” P_4 gibt, dann wählen wir die äußeren Punkte $U = \{P_1, P_2, P_3\}$.

Also ist $VC(C) = 3$.

Beispiel 3.13 (Halbebenen in $\mathbb{R}^d$). ...

Beispiel 3.14 (monotone Disjunktionen). ...

Theorem 3.15. *Für jede endliche Konzeptklasse C gilt $opt(C) \geq VC(C)$.*

Beweis. Sei $VC(C) = k$ und $S = \{x_1, \ldots, x_k\} \subseteq X$ eine Mengen mit $|S| = k$, die von C erschlagen wird. Dann gibt es einen vollständigen Fehlerbaum für C, dessen Knoten auf der Ebene i mit x_i markiert sind. Die Höhe dieses Fehlerbaums ist k, also ist $K(C) = opt(C) \geq k$. □

Auch für diese Ungleichung können wir eine Konzeptklasse C angeben, so daß $opt(C) \geq VC(C)$ eine nur sehr grobe Abschätzung ist.

Beispiel 3.16. Sei C die Konzeptklasse über $X = \{1, \ldots, 2^n\}$, die aus allen Konzepten $c_j = \{1, \ldots, j-1\}$ mit $j = 1, \ldots, 2^n$ besteht. Dann gilt $VC(C) = 1$ aber $opt(C) = n$. (s. Übungen)