AGML, Sommersemester 2020 Vorlesung 15 (5 Seiten)

Komposition

Anne Driemel Letzte Aktualisierung: 18. Juni 2020

Wir haben in der letzten Vorlesung das Boosting kennengelernt, welches schwache Lernalgorithmen miteinander kombiniert um einen starken Lernalgorithmus zu erhalten. Beim Boosting ergibt sich eine neue Hypothesenklasse aus den möglichen Linearkombinationen der Hypothesenklassen der benutzten schwachen Lernalgorithmen. Allerdings erzeugt das Boosting auch eine höhere VC-Dimension und somit die Gefahr, dass Overfitting geschieht. Heute werden wir genauer analysieren, wie sich die Komposition mehrerer Hypothesen auf die VC-Dimension der resultieren Hypothesenklasse auswirkt. Wir betrachten neben dem Boosting auch andere Arten der Komposition.

1 Achsenparallele Hyperquader

Wir schauen uns zunächst die Klasse der Schwellenwertfunktionen in $\mathbb{R}^d$ an und zeigen eine obere Schranke für die VC-Dimension. Schwellwertfunktionen können kombiniert werden, um Hyperquader darzustellen. Dies wird uns als einleitendes Beispiel dienen, bevor wir auf komplexere Kompositionen von Hypothesenklassen eingehen.

Sei die Klasse der Schwellenwertfunktionen in $\mathbb{R}^d$ definiert als Menge von Funktionen der Form $h_{i,a,b} : \mathbb{R}^d \to \{+1, -1\}$ mit $1 \leq i \leq d$, $a \in \mathbb{R}$, $b \in \{+1, -1\}$ und

$$h_{i,a,b}(x_1, \ldots, x_d) = \begin{cases} +b & \text{falls } x_i \geq a \\ -b & \text{sonst} \end{cases}$$

Eine Schwellenwertfunktion $h_{i,a,b}$ entspricht der Partitionierung der Grundmenge durch eine achsenparallelen Hyperebene. Wir definieren die Klasse der Hyperquader in $\mathbb{R}^d$ als Menge von Funktionen $h_{\mathbf{a},\mathbf{b}} : \mathbb{R}^d \to \{+1, -1\}$ definiert durch Vektoren $\mathbf{a} = (a_1, \ldots a_d) \in \mathbb{R}^d$ und $\mathbf{b} = (b_1, \ldots b_d) \in \mathbb{R}^d$ mit $a_i < b_i$ für alle $1 \leq i \leq d$ und

$$h_{\mathbf{a},\mathbf{b}}(x_1, \ldots, x_d) = \begin{cases} +1 & \text{falls } \forall i : a_i \leq x_i \leq b_i \\ -1 & \text{sonst} \end{cases}$$

Es ist leicht zu sehen, dass jeder Hyperquader durch eine Komposition von $2d$ Schwellenwertfunktionen darstellbar ist. Wie können wir nun leicht obere Schranken für die VC-Dimension von Hyperquadern zeigen? Wir analysieren zunächst die VC-Dimension der Schwellenwertfunktion.

Lemma 15.1. *Sei $\mathcal{H}$ die Klasse der Schwellenwertfunktionen mit Grundmenge $\mathbb{R}^d$. $\mathcal{H}$ hat VC-Dimension höchstens* $\max(2 \log_2 d, 8)$.

Beweis. Sei $\mathcal{R}$ das zu $\mathcal{H}$ zugehörige Mengensystem und sei $A \subseteq \mathbb{R}^d$ eine Menge, die von $\mathcal{R}$ aufgespalten wird. Zur Erinnering, das heißt dass für jedes $A' \subseteq A$ eine Menge $r \in \mathcal{R}$ existiert, sodass $A' = r \cap A$. Ziel ist es eine obere Schranke für $|A|$ zu zeigen, denn die VC-Dimension ist definiert als die Kardinalität der größten aufgespaltenen Menge. Dafür sei $t = |A|$.

Wir interessieren uns also für die Anzahl der verschiedenen Mengen $r \cap A$ mit $r \in \mathcal{R}$, also die Größe der Menge $\mathcal{R}|_A$. Gleichzeitig wissen wir, dass es genau 2^t verschiedenen Teilmengen von A gibt, die damit dargestellt werden. Es muss also gelten

$$2^t \leq \left|\mathcal{R}|_A\right|$$

Daraus wollen wir eine obere Schranke für t ableiten.

Die wichtige Beobachtung ist nun, dass es höchstens $2dt$ verschiedene nicht-leere Teilmengen von A gibt, die durch eine achsenparallele Hyperebene abgespalten werden können, da A in jeder Dimension höchstens t verschiedene Koordinaten hat. Das heißt

$$\left|\mathcal{R}|_A\right| \leq dt.$$

Also ist $2^t \leq 2dt$. Nun machen wir eine Fallunterscheidung. Angenommen, dass $t \leq d$. Dann ist $2^t \leq 2d^2$. Durch Logarithmieren auf beiden Seiten ergibt sich $t \leq 2\log_2 2d$. Der zweite Fall ist, dass $t > d$. Daraus ergibt sich analog $t < 2\log_2 2t$. Diese Ungleichung kann für $t \in \mathbb{N}$ nur erfüllt werden wenn $t \leq 8$.

Wir haben also hergeleitet, dass

$$t \leq \max(2\log_2 2d, 8)$$

Da dies für beliebige Mengen A gilt, die durch $\mathcal{R}$ aufgespalten werden, folgt die obere Schranke für die VC-Dimension nun direkt. □

2 Komposition

Definition 15.2 (Komposition)**.** *Sei X eine feste Grundmenge und sei C eine Klasse von Funktionen der Form $f : \{+1, -1\}^k \to \{+1, -1\}$. Sei $\mathcal{H}$ eine Hypothesenklassen mit Grundmenge X und sei $\mathcal{R}$ das zugehörige Mengensystem. Sei $\mathcal{H}_C$ die Hypothesenklasse aller Funktionen $g : X \to \{+1, -1\}$ mit*

$$g(x) = f(h_1(x), \ldots, h_k(x)) \quad \text{und} \quad h_1, \ldots, h_k \in \mathcal{H}, f \in C$$

Wir bezeichnen mit $\mathcal{R}_C$ das zugehörige Mengensystem.

Beispiel 15.3. *Im Fall von Boosting, ist die Klasse C die Menge aller Funktionen der Form $f(y_1, \ldots, y_k) = \operatorname{sign}(\sum_{1 \leq i \leq k} \alpha_i y_i)$ mit $\alpha_i \geq 0$. Der Fakt, dass dies einer Komposition nach Definition 15.2 entspricht, ist dabei unabhängig davon, wie die Gewichte α_i gewählt werden.*

Wir betrachten zunächst den Fall, dass die Klasse C nur aus einer festen Funktion besteht, zum Beispiel der Funktion die in dem zugehörigen Mengensystem die Schnittmenge der positiven Mengen erzeugt:

$$f(y_1, \ldots, y_k) = \begin{cases} +1 & \text{falls } \sum_{i=1}^k y_i = k \\ -1 & \text{sonst} \end{cases} \tag{1}$$

Wir bezeichnen die Komposition in dem Fall einer festen Funktion f mit $\mathcal{H}_f$, beziehungsweise das Mengensystem mit $\mathcal{R}_f$.

Beispiel 15.4. *Sei $\mathcal{H}$ die Klasse der Schwellenwertfunktionen und sei f definiert wie in (1) mit $k = 2d$. Dann ist $\mathcal{R}_f$ die Menge aller Hyperquader in $\mathbb{R}^d$. Das heißt, die Menge enthält alle beschränkten Hyperquader und zusätzlich solche, die in mindestens einer Richtung unbeschränkt sind.*

Beispiel 15.5. *Sei $\mathcal{R}$ das Mengensystem aller Halbräume in $\mathbb{R}^2$ und sei f definiert wie in (1) mit $k = 3$. Dann ist $\mathcal{R}_f$ die Menge aller verallgemeinerten Dreiecke in $\mathbb{R}^2$. Das heißt, die Menge enthält alle beschränkten Dreiecke und zusätzlich solche Dreiecke, die in einer Richtung unbeschränkt sind, siehe Abbildung 1.*

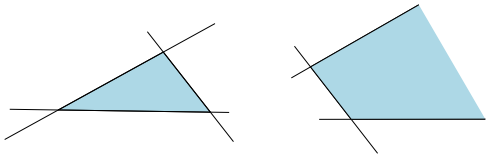

Abbildung 1: Zwei Beispiele von verallgemeinerten Dreiecken.

Wir zeigen nun eine obere Schranke für die VC-Dimension von einfachen Kompositionen, also Kompositionen mit einer festen Funktion f. Dafür zeigen wir erst ein Hilfslemma. Wir notieren mit $\ln x$ den natürlichen Logarithmus zur Basis e.

Lemma 15.6. *Für $x > 0$ und $u \in \mathbb{R}$ gilt $x \leq u \ln x \implies x \leq 2u \ln u$*

Beweis. Wir nutzen, dass für jedes $x > 0$ gilt, dass $\ln x \leq \sqrt{x}$.

$$\begin{aligned} & x \leq u \ln x \\ \implies \quad & x \leq u\sqrt{x} \\ \implies \quad & \ln x \leq \ln u + \tfrac{1}{2} \ln x \\ \implies \quad & \tfrac{1}{2} \ln x \leq \ln u \\ \implies \quad & \ln x \leq 2 \ln u \end{aligned}$$

Die Aussage folgt nun durch einfaches Einsetzen. □

Satz 15.7. *Sei $\mathcal{H}$ eine Hypothesenklasse mit Grundmenge X und VC-Dimension höchstens d mit $3 \leq \mathrm{d} < \infty$. Sei $f : \{+1, -1\}^k \to \{+1, -1\}$ eine feste Funktion mit $k \geq 3$. Die VC-Dimension der Komposition $\mathcal{H}_f$ ist höchstens $4\mathrm{d}k \ln(2\mathrm{d}k)$.*

Beweis. Sei $A \subseteq X$ eine Menge, die von dem zugehörigen Mengensystem $\mathcal{R}_f$ aufgespalten wird. Wir folgen nun derselben Strategie wie in dem Beweis zu Lemma 15.1. Die Herausforderung besteht darin, eine obere Schranke für $\left|\mathcal{R}_f|_A\right|$ zu finden. Zur Erinnerung, diese Menge ist wie folgt definiert.

$$\mathcal{R}_f|_A = \{\ r \cap A \ \mid \ r \in \mathcal{R}_f\ \}$$

Laut Definition des Mengensystems wissen wir, dass für jede Menge $r \in \mathcal{R}_f$ Hypothesen $h_1, \ldots, h_k \in \mathcal{H}$ existieren, sodass

$$r = \{\ x \in X \ \mid \ f(h_1(x), \ldots, h_k(x)) = 1\ \}$$

Also ist

$$r \cap A = \{\ x \in A \ \mid \ f(h_1|_A(x), \ldots, h_k|_A(x)) = 1\ \}$$

Daraus folgt, dass die Anzahl der verschiedenen Mengen $r \cap A$ mit $r \in \mathcal{R}_f$ nur von Funktionen in $\mathcal{H}|_A$ abhängt. Deren Anzahl ist durch die Wachstumsfunktion $\Pi_{\mathcal{H}}(t)$ beschränkt. Insbesondere entsteht eine Menge $r \cap A$ indem wir k Hypothesen aus $\mathcal{H}|_A$ auswählen. Also ist laut dem Wachstumslemma

$$\left|\mathcal{R}_f|_A\right| \leq \left|\mathcal{H}|_A\right|^k \leq (\Pi_{\mathcal{H}}(t))^k \leq \left(\frac{\mathrm{e}t}{\mathrm{d}}\right)^{\mathrm{d}k} \leq t^{\mathrm{d}k} \tag{2}$$

wobei wir nutzen, dass $\mathrm{d} \geq 3$ angenommen wird.

Daraus leiten wir ab, dass $2^t \leq t^{\mathrm{d}k}$ und durch Logarithmieren mit dem natürlichen Logarithmus auf beiden Seiten ergibt sich

$$t \ln 2 \leq (\mathrm{d}k) \ln t$$

Da $\ln 2 > 0.5$ ergibt sich durch Umformen $t \leq 2\mathrm{d}k \ln t$. Nun können wir Lemma 15.6 anwenden und erhalten

$$t \leq 4\mathrm{d}k \ln(2\mathrm{d}k)$$

Da dies für beliebige Mengen A gilt, die durch das Mengensystem aufgespalten werden, ergibt sich die obere Schranke für die VC-Dimension. □

Aus obigen Satz folgt nun für die Mengensysteme in unseren Beispielen, dass die VC-Dimension von Dreiecken durch eine Konstante beschränkt ist und für die Hyperquader in $\mathbb{R}^d$ ergibt sich zusammen mit Lemma 15.1 eine obere Schranke von $O(d \log^2 d)$.

3 VC-Dimension des Boostings

Satz 15.8. *Sei $\mathcal{H}$ eine Hypothesenklasse mit Grundmenge X und VC-Dimension höchstens* d *mit* $3 \leq \mathrm{d} < \infty$. *Sei C die Klasse von Funktionen $f : \{+1, -1\}^k \to \{+1, -1\}$ der Form $f(y_1, \ldots, y_k) = \operatorname{sign}(\sum_{1 \leq i \leq k} \alpha_i y_i)$ mit $\alpha_i \geq 0$ und sei $k \geq 3$. Die VC-Dimension der Komposition $\mathcal{H}_C$ ist höchstens* $4(\mathrm{d}+1)k \ln(2(\mathrm{d}+1)k)$.

Beweis. Wir folgen wieder derselben Strategie wie in dem Beweis zu Lemma 15.1. Der Beweis ist ähnlich zu dem Beweis zu Satz 15.7. Wir müssen allerdings zusätzlich die verschiedenen Funktionen in C beachten.

Sei $A \subseteq X$ eine Menge, die von $\mathcal{R}_C$ aufgespalten wird und sei $t = \big|A\big|$. Wie zuvor wollen wir wieder eine obere Schranke für die Anzahl der verschiedenen Mengen in $\mathcal{R}_C|_A$ finden, und nutzen, dass $2^t \leq \big|\mathcal{R}_C|_A\big|$ gelten muss. Zur Erinnerung,

$$\mathcal{R}_C|_A = \{\ r \cap A \ \mid \ r \in \mathcal{R}_C\ \}$$

Betrachte eine konkrete Teilmenge $A' \subseteq A$. Falls A' abgespalten wird, dann existiert eine Menge $r \in \mathcal{R}_C$ sodass $A' = r \cap A$. Die Menge r ist definiert durch konkrete Hypothesen $h_1, \ldots, h_k \in \mathcal{H}$ und eine konkrete Funktion $f \in C$ mit

$$r = \{\ x \in X \ \mid \ f(h_1(x), \ldots, h_k(x)) = 1\ \}$$

Wie zuvor haben wir

$$r \cap A = \{\ x \in A \ \mid \ f(h_1|_A(x), \ldots, h_k|_A(x)) = 1\ \}$$

Wir wissen aus der vorherigen Analyse im Beweis zu Satz 15.7, dass für ein festes $f \in C$ höchstens $(\Pi_{\mathcal{H}}(t))^k$ verschiedene Mengen erzeugt werden können, weil wir uns auf die Funktionen in $\mathcal{H}|_A$ beschränken können.

Ähnlich wollen wir nun auch die Funktionen $f \in C$ beschränken. Dafür stellen wir zunächst eine andere Frage. Wieviele Mengen können erzeugt werden, wenn wir k Hypothesen aus $\mathcal{H}$ festhalten und $f \in C$ frei wählen können?

Seien $h_1, \ldots, h_k$ fest und betrachte die Menge

$$B = \{\ (h_1(x), \ldots, h_k(x)) \ \mid \ x \in A\ \}$$

Beachte, dass $\big|B\big| = \big|A\big| = t$.

Wir betrachten nun das Mengensystem $\mathcal{R}'$ mit Grundmenge $\{+1,-1\}^k$ in der jede Menge definiert ist durch eine Funktion $f \in C$ mit

$$r_f = \left\{ (y_1,\dots,y_k) \in \{+1,-1\}^k \;\middle|\; f(y_1,\dots,y_k) = 1 \right\}$$

Betrachten wir dieses Mengensystem genauer, dann stellen wir fest, dass es sich um ein Mengensystem von Halbräumen in $\mathbb{R}^k$, beschränkt auf die Grundmenge $\{+1,-1\}^k$, handelt.

Insbesondere ist f definiert durch $\alpha_1,\dots,\alpha_k \in \mathbb{R}$ mit

$$f(y_1,\dots,y_k) = \begin{cases} 1 & \text{falls } \sum_{1\le i\le k} \alpha_i y_i \ge 0 \\ -1 & \text{sonst} \end{cases}$$

Für $\mathbf{w} = (\alpha_1,\dots,\alpha_k)$ und $u = 0$, sowie $\mathbf{y} = (y_1,\dots,y_k)$ ist also

$$\mathbf{y} \in r_f \quad \Leftrightarrow \quad \langle \mathbf{w}, \mathbf{y} \rangle \ge u$$

Das heißt, r_f enthält genau solche $\mathbf{y} \in \{+1,-1\}^k$ die in dem Halbraum liegen, der durch $\mathbf{w}$ und u definiert ist. Da die VC-Dimension von Halbräumen in $\mathbb{R}^k$ gleich k ist, erhalten wir mit dem Wachstumslemma

$$\big|\mathcal{R}'|_B\big| \le \Pi_{\mathcal{R}'}(t) \le \left(\frac{\mathrm{e}t}{k}\right)^k$$

Diese Erkenntnis können wir nun verwenden um eine obere Schranke für die Anzahl der verschiedenen Mengen $r \cap A$ mit $r \in \mathcal{R}_C$ herzuleiten. Indem wir k verschiedene Hypothesen aus $\mathcal{H}$ auswählen, können wir höchstens $(\Pi_{\mathcal{H}}(t))^k$ verschiedene Mengen B erzeugen. Jede solche Menge B entspricht einer Art, den Elementen in A jeweils k Labels aus $\{+1,-1\}$ zuzuweisen. Nun können wir für jede solche Menge B eine Funktion f auswählen. Für eine feste Menge B können wir dadurch höchstens $\Pi_{\mathcal{R}'}(t)$ verschiedene Mengen erzeugen. Also erhalten wir

$$\big|\mathcal{R}_C|_A\big| \le (\Pi_{\mathcal{H}}(t))^k \,\Pi_{\mathcal{R}'}(t) \le \left(\frac{\mathrm{e}t}{\mathrm{d}}\right)^{\mathrm{d}k} \left(\frac{\mathrm{e}t}{k}\right)^k \le t^{(\mathrm{d}+1)k} \tag{3}$$

wobei wir nutzen, dass $k \ge 3 \ge \mathrm{e}$ und $\mathrm{d} \ge 3 \ge \mathrm{e}$. Nun können wir wieder Lemma 15.6 benutzen und erhalten

$$t \le 4(\mathrm{d}+1)k \ln(2(\mathrm{d}+1)k)$$

□

Referenzen

- Understanding Machine Learning, Kapitel 10.3 (VC-Dimension of Boosting)
- Foundations of Machine Learning, Kapitel 7.3 (VC-Dimension of Boosting)