Modulkonzept zur VC Generalization Bound

Förderkennzeichen:	01IS17073
Vorhabenbezeichnung:	Verbundprojekt: COSY-Entwicklung von sieben Praxis-Versuchen zum Thema Datenanalyse und Maschinelles Lernen an zwei Hochschulstandorten
Zugehörigkeit:	Modulkonzept Versuch 4 - Umweltdaten

Allgemeines:

Die VC-Dimension (nach der Vapnik-Chervonenkis-Theorie) ist ein Maß für die Dimension eines Raumes von Funktionen, die durch einen statistischen Klassifizierer gelernt werden können. Es ist definiert als die Kardinalität der maximalen Anzahl von Punkten, die der Algorithmus trennen kann. Graphisch veranschaulichen lässt sich das anhand eines Klassifizierers $f : x \to \{0, 1\}$ mit $x = \mathbb{R}^2$. Die Samples können als Punkte in der Ebene dargestellt werden. Wir können dann zeigen, dass in 2D alle Sets aus drei Punkten von einer Geraden (linearen Funktion) getrennt werden können (siehe Abbildung).

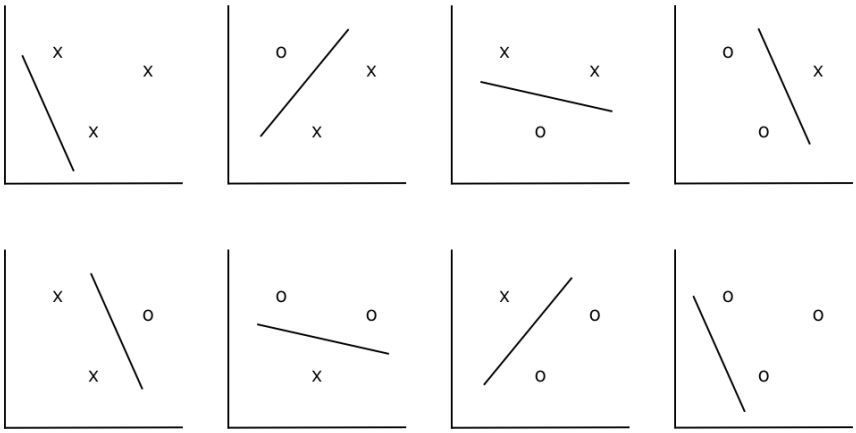

Abbildung 1: Trennbarkeit aller Sets aus drei Punkten in 2D durch eine lineare Funktion

Darüber hinaus kann gezeigt werden, dass dies mit vier Punkten nicht mehr möglich ist (siehe Abbildung).

Daraus kann ein Grenzwert für den Fehler ϵ, mit dem ein unbekanntes System mit einer Wachstumsfunktion $m_{\mathcal{H}}$ und einer Wahrscheinlichkeit δ aus N Samples gelernt werden kann, bestimmt werden. Der In-Sample-Error E_{in} bezeichnet dabei den Fehler auf den Trainingsdaten und der Out-of-Sample-Error E_{out} den Fehler auf den Testdaten. Dieser, durch $E_{in}(h) - E_{out}(h)$ gegebene,

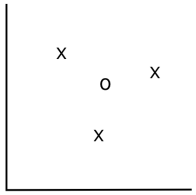

Abbildung 2: Nicht-Trennbarkeit aller Sets aus vier Punkten in 2D durch eine lineare Funktion

Grenzwert ist die VC Generalization Bound und ist durch folgendes Theorem gegeben:

$$\epsilon \geq \sqrt{\frac{8}{N} \ln \frac{4m_{\mathcal{H}}(2N)}{\delta}} \quad \text{mit} \quad \epsilon = E_{in}(h) - E_{out}(h)$$

Da die Wachstumsfunktion $m_{\mathcal{H}}$ die polynomielle Obergrenze $N^{d_{vc}} + 1$ hat, wobei d_{vc} die VC-Dimension des gelernten System ist, kann sie durch diese ersetzt werden.

Die VC Generalization Bound kann dazu verwendet werden die Anzahl an Samples N zu bestimmen mit denen eine Hypothese mit einer VC Dimension d_{vc} mit einer Wahrscheinlichkeit δ mit einem maximalen Fehler ϵ gelernt werden kann. Dazu wird folgende Bound genutzt:

$$N \geq \frac{8}{\epsilon^2} \ln \frac{4(2N)^{d_{vc}} + 1}{\delta}$$

N wird dabei iterativ bestimmt indem zuerst ein geschätzter Wert eingesetzt wird der dann so lange durch das Ergebnis ersetzt wird bis er konvergiert.

In der Praxis ist jedoch meistens ein Datensatz mit fester Größe gegeben, so dass N ebenfalls fest ist. In diesem Fall kann ein Grenzwert für die Performance mit diesem Datensatz bestimmt werden. Dazu wird folgende Bound verwendet um mit einer Wahrscheinlichkeit von $1 - \delta$ den out-of-sample error E_{out} zu berechnen:

$$E_{out}(h) \leq E_{in}(h) + \sqrt{\frac{8}{N} \ln \frac{4((2N)^{d_{vc}} + 1)}{\delta}}$$

Die VC Generalization Bound ist universell, da sie für alle Hypothesenmengen, Lernalgorithmen, Eingabewertebereiche, Wahrscheinlichkeitsverteilungen und binäre Zielfunktionen gilt. Darüber hinaus lässt sie sich auch auf nicht-binäre Zielfunktionen erweitern.

Aufgaben:

- Bestimmung der benötigten Samples zum Lernen einer Hypothese mit zwei Dimensionen und einem maximalen Fehler von 0.05 mit einer Wahrscheinlichkeit von 10%.
- Bestimmung des in-sample Fehlers und des out-of-sample Fehlers für einen gegebenen Datensatz.

- Bestimmung der durch die VC Bound gegebenen Schranke für die Performance für den Datensatz und Vergleich mit dem tatsächlichen Fehler.

Ziele:

- Die Studierenden verstehen den Zusammenhang zwischen VC-Dimension, der Anzahl der Samples und der Performance eines Lernmodells
- Die Studierenden verstehen die Aussage der VC Theory und die Bedeutung für praktische Probleme

Literatur:

- Abu-Mostafa, Y. S., Magdon-Ismail, M., & Lin, H. (2012). Learning from data: a short course (S. 39-75). [United States]: AMLBook.com