Konzeptaktivierungsvektor (CAV)

Autor und Herausgeber

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland (ORCID 0000-0001-7538-0990)
Editors: Konstantina Olioumtsevits (ORCID 0000-0003-2697-8659), Ekkehard Schnoor (ORCID 0000-0001-8252-1499), Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Zusammenfassung

Ein Konzeptaktivierungsvektor (CAV) ist eine Richtung in den Aktivierungen einer verborgenen Schicht eines tiefen Netzes, die für ein von einer Benutzerin oder einem Benutzer benanntes Konzept steht. Es werden Datenpunkte bereitgestellt, die das Konzept aufweisen, sowie Datenpunkte, bei denen dies nicht der Fall ist, und ein linearer Klassifikator wird darauf trainiert, die beiden anhand ihrer Aktivierungen zu unterscheiden. Der CAV ist die Richtung senkrecht zu der Hyperebene, die sie trennt, und zeigt zur Seite des Konzepts. Verschiebt man die Aktivierungen eines Datenpunkts ein wenig entlang des CAV, ändert sich der Wert, den das tiefe Netz einem Label zuweist; die Änderungsrate ist die Sensitivität der Vorhersage gegenüber dem Konzept. Der Anteil der Datenpunkte mit einer positiven Sensitivität ist der TCAV-Wert. Dieser Wert gehört zu einer Schicht und einem Label; er aggregiert die Sensitivitäten pro Datenpunkt und beschreibt dadurch eine ganze Klasse von Datenpunkte statt einer einzelnen Vorhersage.

Definition

Ein tiefes Netz sagt aus dem Wetter von fünf aufeinanderfolgenden Tagen an einer Station vorher, ob es am sechsten Tag regnet. Ob es den Begriff einer Hitzewelle, also fünf heiße Tage in Folge, nutzt, um zu dieser Vorhersage zu gelangen, lässt sich nicht durch die Betrachtung eines einzelnen Neurons entscheiden. Eine Hitzewelle ist kein Merkmal, mit dem das Netz konstruiert wurde, da seine Merkmale die täglichen Temperaturen und Regenmengen sind. Ebenso wenig muss sie durch ein einzelnes Neuron repräsentiert werden (Kim et al., 2018, Abschn. 2.3). Ein Konzept ist eine Eigenschaft, die man benennen und durch Beispiele belegen kann, für deren Berechnung das Netz jedoch nicht konstruiert wurde, wie etwa eine Hitzewelle. Das Konzept wird stattdessen durch Beispiele spezifiziert: Fünftagesfenster, die eine Hitzewelle bilden, als positive Beispiele, und Fenster, die keine bilden, als negative Beispiele (Kim et al., 2018, Abschn. 3.1). Ein CAV übersetzt ein solches Konzept in eine Richtung in den Aktivierungen einer verborgenen Schicht – die Richtung, entlang derer die positiven Beispiele getrennt von den negativen liegen.

Betrachtet sei ein solches tiefes Netz mit mehreren verborgenen Schichten, das darauf trainiert ist, das Label eines Datenpunkts aus seinem Merkmalsvek-

tor vorherzusagen; das trainierte Netz ist die gelernte Hypothese $\widehat{h}$. Man hält eine verborgene Schicht fest und schreibt $\mathbf{z} = f(\mathbf{x})$ für ihre Aktivierungen, wobei f die von den Schichten bis zu diesem Punkt berechnete Funktion ist. Das tiefe Netz wird auf die positiven und auf die negativen Beispiele eines Konzepts $\mathcal{C}$ angewandt. Ein binärer linearer Klassifikator $g(\mathbf{z})$, dessen Merkmale diese Aktivierungen sind, wird darauf trainiert, die beiden zu unterscheiden. Seine Entscheidungsgrenze ist eine Hyperebene; der CAV für $\mathcal{C}$ ist der Normalenvektor $\widehat{\mathbf{w}}$ dieser Hyperebene (Kim et al., 2018, Abschn. 3.2).

Eine Hyperebene besitzt zwei Normalenvektoren, von denen einer in jede Seite zeigt. Der CAV ist derjenige, der auf die Seite zeigt, die $g(\mathbf{z})$ der Konzeptklasse zuweist. Das Training von $g(\mathbf{z})$ mit den positiven Beispielen als positiver Klasse liefert diese Orientierung, und wenn die beiden Mengen von Aktivierungen linear trennbar sind, ist es die Seite der positiven Beispiele. In der Terminologie der mechanistischen Interpretierbarkeit definiert diese Richtung ein Merkmal des trainierten tiefen Netzes, nämlich die Projektion der Aktivierungen auf sie: ein für den Menschen interpretierbares Konzept, repräsentiert durch eine Richtung im Raum der Aktivierungen.

Abb. 1 zeigt den CAV des Hitzewellen-Konzepts für ein tiefes Netz mit zwei verborgenen Schichten von je 16 Neuronen, das auf dem täglichen Wetter in Krems an der Donau der Trainingsjahre 2000–2018 trainiert und auf den Jahren 2019–2024 getestet wurde. Ein Datenpunkt ist ein Fenster von fünf Tagen, sein Merkmalsvektor enthält die tägliche Höchst- und Tiefsttemperatur sowie die Regenmenge jedes Tages, und sein Label gibt an, ob am sechsten Tag mindestens 1 mm Regen fiel. Das Konzept wird durch 40 Hitzewellen spezifiziert – Fenster mit einem täglichen Maximum von mindestens 30 °C an allen fünf Tagen – sowie 40 Fenster, auf die dies nicht zutrifft, gezogen aus den Fenstern der Aufzeichnung. Die eingezeichnete Ebene verläuft durch die mittlere Aktivierung der ersten verborgenen Schicht und wird vom CAV sowie der zu ihm orthogonalen Richtung aufgespannt, entlang derer die Aktivierungen am stärksten variieren. Die Konzept-Hyperebene schneidet die Ebene daher in einer vertikalen Geraden, und der CAV bildet ihre horizontale Achse.

In der Abbildung treten zwei Klassifikationsprobleme auf. Das erste ist die Konzept-Klassifikation, die $g(\mathbf{z})$ auf den Aktivierungen löst. Das zweite ist die Aufgabe, für die das tiefe Netz trainiert wurde und deren Entscheidungsgrenze in der Ebene die durchgezogene Linie ist: Rechts von ihr, wo jede Hitzewelle liegt, wird Regen vorhergesagt. Der CAV hängt von den Beispielen ab, die das Konzept spezifizieren: Ein Training von $g(\mathbf{z})$ auf allen 9127 Fenstern der Aufzeichnung dreht ihn stattdessen um 26 Grad.

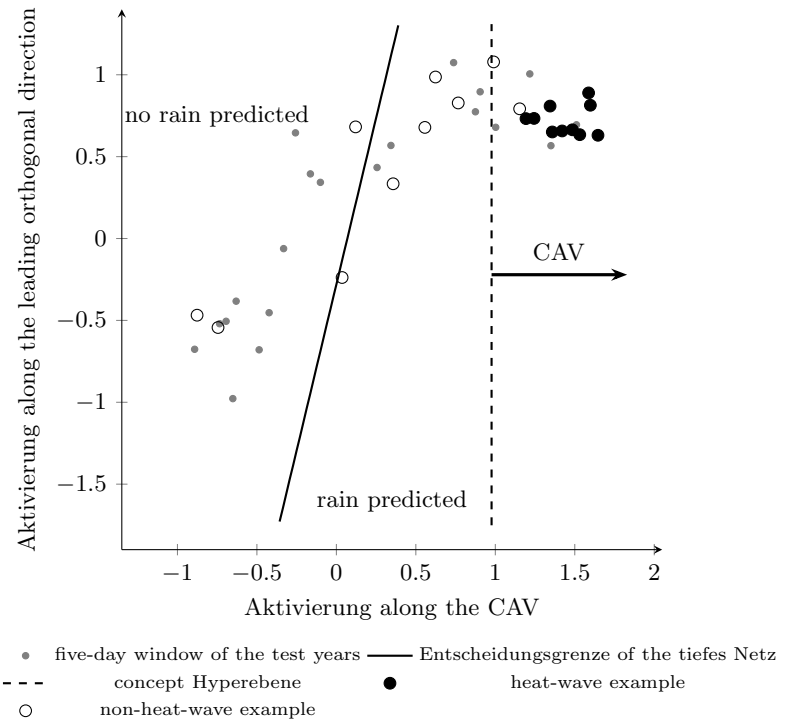


Abbildung 1: Der Hitzewellen-CAV eines tiefen Netzes, das Regen in Krems an der Donau aus den fünf vorangegangenen Tagen vorhersagt, dargestellt in der Ebene durch die mittlere Aktivierung seiner ersten verborgenen Schicht, aufgespannt vom CAV (horizontale Achse) und der zu ihm orthogonalen Richtung, entlang derer die Aktivierungen am stärksten variieren. Graue Punkte sind 20 der 400 zufällig gezogenen Fünftagesfenster der Testjahre. Ausgefüllte Kreise sind 10 der 40 als Hitzewellen-Beispiele übergebenen Fenster und offene Kreise 10 der 40 als Nicht-Hitzewellen-Beispiele übergebenen Fenster, jeweils jedes vierte; die gestrichelte Linie ist die Entscheidungsgrenze des auf allen 80 trainierten linearen Klassifikators, und der CAV ist ihr Normalenvektor. Die durchgezogene Linie ist die Entscheidungsgrenze des tiefen Netzes in dieser Ebene, wobei rechts von ihr Regen vorhergesagt wird. Daten erzeugt mit `pythondemos/cav.py`

Wie jede Erklärung muss ein CAV verständlich und originalgetreu sein (siehe Erklärung). Verständlich ist er konstruktionsbedingt: Die Beispiele, die das Konzept definieren, werden vorgegeben. Die Originaltreue ist eine gesonderte Frage, da das Konzept eine Richtung in den Aktivierungen darstellt, unabhängig davon, ob das tiefe Netz darauf anspricht oder nicht. Das Testen mit CAVs (TCAV) entscheidet darüber.

Man schreibt das trainierte tiefe Netz als Verkettung zweier Funktionen: den Aktivierungen $\mathbf{z} = f(\mathbf{x})$ der gewählten verborgenen Schicht und einem Wert

$s(\mathbf{z})$, den die verbleibenden Schichten einem Label zuweisen. Die konzeptionelle Sensitivität eines Datenpunkts ist die Richtungs-Ableitung dieses Werts entlang des normierten CAV $\mathbf{v} := \widehat{\mathbf{w}}/\|\widehat{\mathbf{w}}\|$, was für ein differenzierbares s dem Skalarprodukt des Gradienten von s mit $\mathbf{v}$ entspricht (Rudin, 1976, Beispiel 9.18),

$$S(\mathbf{x}) := \langle \nabla s\big(f(\mathbf{x})\big), \mathbf{v} \rangle. \tag{1}$$

Sie misst, wie stark sich der Wert für dieses Label verändert, wenn sich die Aktivierungen ein wenig in Richtung des Konzepts bewegen (Kim et al., 2018, Abschn. 3.3). Die Konvention für die Orientierung verleiht dem Vorzeichen seine Bedeutung: $S(\mathbf{x}) > 0$ besagt, dass der Wert in Richtung des Konzepts ansteigt, und der entgegengesetzte Normalenvektor würde jede Sensitivität umkehren.

Die Sensitivität (1) liefert eine Zahl pro Datenpunkt. TCAV fasst diese Zahlen über ein Label zusammen: Der TCAV-Wert ist der Anteil der Datenpunkte mit diesem Label, für die $S(\mathbf{x}) > 0$ gilt (Kim et al., 2018, Abschn. 3.4). Eine zufällige Menge von Nicht-Konzept-Datenpunkte liefert ebenfalls einen CAV, sodass ein einzelner solcher Anteil allein noch kein Beleg ist. Ein Zufallskonzept ist eine Menge zufällig gezogener Datenpunkte, die so mit einem Label versehen werden, als trügen sie ein Konzept. Es erhöht den Wert mit derselben Wahrscheinlichkeit, mit der es ihn senkt, sodass sein TCAV-Wert, der Anteil positiver Sensitivitäten, im Mittel ein halb beträgt. Der CAV wird daher wiederholt gegen viele neue Ziehungen von Nicht-Konzept-Datenpunkte angepasst, und die resultierenden TCAV-Werte werden mit jenen von Zufallskonzepten verglichen: Ein Konzept, auf das das tiefe Netz anspricht, erzielt beständig Werte abseits von ein halb, ein zufälliges hingegen nicht (Kim et al., 2018, Abschn. 3.5).

Die gelieferte Erklärung ist eine einzelne Zahl, die einem Konzept, einem Label und einer Schicht zugeordnet ist. Im Gegensatz zur Sensitivität $S(\mathbf{x})$, die für einen einzelnen Datenpunkt definiert ist, beschreibt sie eine ganze Klasse von Datenpunkte statt einer einzelnen Vorhersage (Kim et al., 2018, Abschn. 2.1).

Für das tiefe Netz aus Abb. 1 erhöht eine Bewegung entlang des Hitzewellen-CAV den Regenwert bei jedem Testfenster, auf das Regen folgt. Eine Kältewelle besteht aus fünf Tagen mit einem Minimum von höchstens $-5\,^{\circ}\mathrm{C}$, und eine Bewegung entlang ihres CAV senkt diesen Wert bei jedem dieser Fenster. Die Aufzeichnungen stimmen damit überein: Regen folgt auf 30 % der Hitzewellen und 16 % der Kältewellen, gegenüber 23 % über alle Fenster.

Sowohl das Training des linearen Klassifikators $g(\mathbf{z})$ als auch die Auswertung der Sensitivität $S(\mathbf{x})$ lesen Größen im Inneren des tiefen Netzes aus: die Aktivierungen $f(\mathbf{x})$ der gewählten Schicht und den Gradienten $\nabla s(\mathbf{z})$ des Werts nach ihnen. Ein CAV ist daher nicht verfügbar für ein System, das lediglich Vorhersagen für übergebene Merkmalsvektoren und sonst nichts zurückgibt, wie etwa ein tiefes Netz, das hinter einer Schnittstelle betrieben oder von einem Anbieter bereitgestellt wird. Eine durch lokale interpretierbare modellunabhängige Erklärungen (LIME) erzeugte Erklärung benötigt im Gegensatz dazu ausschließlich Vorhersagen: Ihre lokale Approximation wird an Vorhersagen angepasst, die durch Abfragen der gelernten Hypothese $\widehat{h}$ gewonnen werden (Ribeiro et al., 2016, Abschn. 3). Eine zweite Konsequenz aus dem Auslesen der internen

Größen ist, dass ein CAV keine Eigenschaft von $\widehat{h}$ allein ist: Die Aufteilung in f und s hängt von der Wahl der Schicht ab, und zwei tiefe Netze, die dieselbe $\widehat{h}$ über unterschiedliche Schichten realisieren, können unterschiedliche CAVs aufweisen.

Beide Funktionen, die ein CAV einbezieht, sind auf dem Raum der Aktivierungen definiert: $g(\mathbf{z})$ approximiert dort die Konzeptzugehörigkeit, und die Sensitivität $S(\mathbf{x})$ ist eine Richtungs-Ableitung des Werts s auf demselben Raum. Ein CAV betrifft daher die Interpretierbarkeit des tiefen Netzes und nicht seine Erklärbarkeit; die mechanistische Interpretierbarkeit untersucht solche Richtungen.
Siehe auch: tiefes Netz, Merkmal, linearer Klassifikator, vertrauenswürdige künstliche Intelligenz (vertrauenswürdige KI), Interpretierbarkeit, mechanistische Interpretierbarkeit, Transparenz, Erklärung, Erklärbarkeit, Neuron, Entscheidungsgrenze, Hyperebene, LIME.

Querverweise

Literatur

1. Ribeiro et al. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, pp. 1135–1144.

2. Rudin (1976). Principles of Mathematical Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.

3. Kim et al. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Proc. 35th Int. Conf. Mach. Learn., pp. 2668–2677.