Membership-Inference-Angriff

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 Membership-Inference-Angriff zielt darauf ab festzustellen, ob ein bestimmter Datenpunkt zum Trainieren eines machine learning (ML)-Modells verwendet wurde. Die Grundidee ist, dass eine gelernte Hypothese genauere Vorhersagen für Datenpunkte liefert, die denen ähneln, die für das Training verwendet wurden. Ein Angreifer fragt die gelernte Hypothese ab und vergleicht den Verlust ihrer Vorhersage mit einem Schwellenwert. Minderungsstrategien umfassen die Regularisierung, die eine ML-Methode über das Trainingsdatensatz hinaus verallgemeinern lässt: Die gelernte Hypothese verhält sich dann für Datenpunkte im Trainingsdatensatz und für solche außerhalb davon ähnlich, was es einem Angreifer erschwert, die Datenpunkte des Trainingsdatensatzs zu erraten.

Definition

Ein freiwilliger Beobachter in Krems an der Donau (Österreich) meldet die Lufttemperatur um 06:00 (UTC) und um 15:00 (UTC) für ausgewählte Tage. Jeder Tag ist ein Datenpunkt $\mathbf{z} = (x, y)$, dessen Merkmal x der 06:00-Messwert dieses Tages und dessen Label y der 15:00-Messwert dieses Tages ist, und die gemeldeten Tage bilden das Trainingsdatensatz $\mathcal{D}^{(\text{train})}$. Ein Unternehmen lernt eine Hypothese $\widehat{h}$ aus $\mathcal{D}^{(\text{train})}$, um den 15:00-Messwert eines Tages aus seinem 06:00-Messwert vorherzusagen. Das Unternehmen veröffentlicht $\widehat{h}$, sodass jeder sie bei jedem 06:00-Messwert abfragen kann. Die GeoSphere-Wetterstation in Krems veröffentlicht die Lufttemperatur jeder Stunde jedes Tages. Jeder, der beides hat, kann versuchen festzustellen, ob der Bericht eines bestimmten Tages für das Training verwendet wurde. Da die Berichte von einer Person stammen, besagt ein Ja, dass der Beobachter an diesem Tag in der Nähe von Krems an der Donau war. Ob der Beobachter an einem bestimmten Tag in Krems war, sind personal data, auch wenn eine Temperatur dies nicht ist (European Parliament and Council of the European Union, 2016, Art. 4). Dieses Szenario ist ein Beispiel für einen Membership-Inference-attack: ein privacy attack, bei dem ein Angreifer versucht festzustellen, ob ein bestimmter Datenpunkt zum Trainieren eines machine learning (ML)-Modells verwendet wurde (Shokri et al., 2017; Vassilev et al., 2024, Sect. 2.4.2).

Für einen bestimmten Tag liest der Angreifer das Merkmal x und das Label y dieses Tages aus den Aufzeichnungen der Station, berechnet die Vorhersage $\widehat{h}(x)$

und vergleicht sie mit y. Ein Verlust $\left(y-\widehat{h}(x)\right)^2$ unterhalb eines Schwellenwerts zeigt an, dass der Bericht dieses Tages für das Training verwendet wurde.

Abb. 1 führt diesen attack an den 31 Tagen des Oktobers 2024 in Krems aus. Ein Schwellenwert kennzeichnet alle 31 Tage korrekt: Der Verlust ist an jedem Tag, den der Beobachter gemeldet hat, null und an keinem anderen Tag unter $0.24\,^\circ\mathrm{C}^2$. Die Tage, die es als gemeldet deklariert, bilden zwei Blöcke aufeinanderfolgender Tage, den 7. bis 13. und den 20. bis 26. Oktober, was die Zeit ist, in der der Beobachter dort war. Die Hypothese hinter der Abbildung wird durch lineare Regression auf den Merkmale $x'_j = \cos\left((j-1)\pi u\right)$ für $j = 1,\ldots,d$ erhalten, wobei u der 06:00-Messwert x ist, der über den Monat auf $[0,1]$ skaliert wurde; für $d = 14$, ein Merkmal pro gemeldetem Tag, reproduziert least squares jeden gemeldeten Tag exakt.

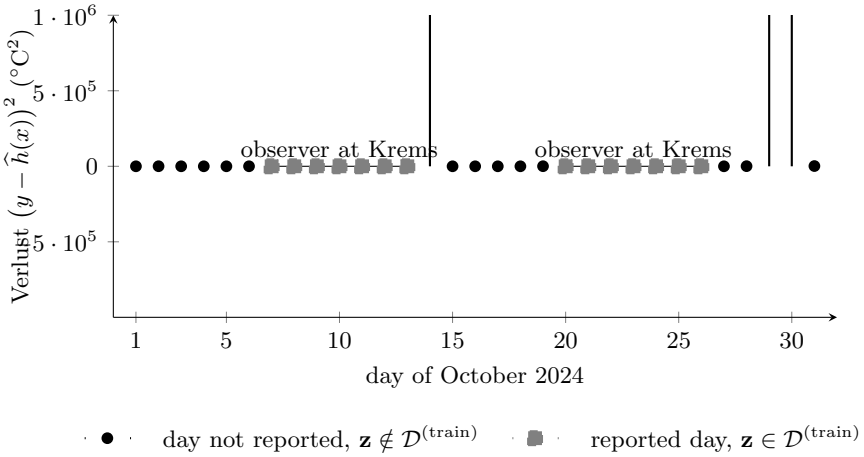

Abbildung 1: Der Verlust, den die veröffentlichte Hypothese an jedem Tag des Oktobers 2024 in Krems an der Donau verursacht, berechnet von einem Angreifer aus den eigenen Aufzeichnungen der Wetterstation. Jeder Tag, den der Beobachter nicht gemeldet hat, trägt einen durchgezogenen Stiel, der von Null bis zu seinem Verlust gezeichnet und von einer Scheibe gekrönt ist. Die 14 Tage, die der Beobachter gemeldet hat, tragen überhaupt keinen Stiel: Die auf der horizontalen Achse liegenden Quadrate markieren einen Verlust von Null. Die Tage, die diese Quadrate markieren, sind der 7. bis 13. und der 20. bis 26. Oktober, die beiden Aufenthalte, sodass die Verluste nicht nur aussagen, welche Tage für das Training verwendet wurden, sondern auch, wann der Beobachter in Krems war. Daten generiert von `pythondemos/membershipinferenceattack.py`

Die Zahl d legt die effektive Dimension des zugrunde liegenden Hypothesenraums fest und damit, wie stark die ML-Methode überanpassen kann. Abb. 2 zeigt, wie viel der attack daraus gewinnt. Man fixiert einen Schwellenwert τ auf den Verlust. Von den 14 Tagen, die der Beobachter gemeldet hat, ist der Anteil, der einen Verlust von höchstens τ aufweist, die Erkennungsrate. Von den 17

Tagen des Monats, die der Beobachter nicht gemeldet hat, ist der Anteil, der einen Verlust von höchstens τ aufweist, die Falschalarmrate. Der Vorteil des attacks ist der größte Wert der Erkennungsrate minus der Falschalarmrate über alle Schwellenwerte hinweg. Ein Angreifer, der an jedem Tag mit Ja oder an jedem Tag mit Nein antwortet, lässt die beiden Raten gleich und den Vorteil null, weshalb der Vorteil misst, was der attack gegenüber dem Raten gewinnt, und ein Schwellenwert, der die beiden Gruppen trennt, macht ihn zu eins. Der Schwellenwert wird mit bereits bekannter Antwort gewählt, sodass der Vorteil das Maximum ist, das jeder Schwellenwert erreichen könnte, und nicht das, was ein Angreifer erzielen würde. Mit zwei Merkmale beträgt er bereits 0.43, sodass selbst der kleinste Hypothesenraum hier etwas preisgibt, und er erreicht eins, sobald d die Anzahl der gemeldeten Tage erreicht.

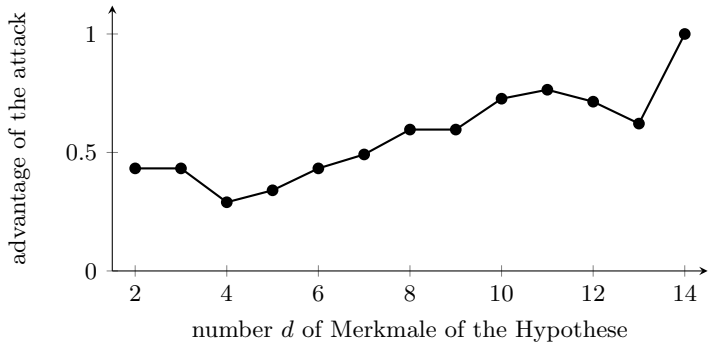

Abbildung 2: Wie sich der attack von Abb. 1 verbessert, wenn der Hypothesenraum wächst. Die Kurve zeigt den Vorteil, die größte Differenz über alle Schwellenwerte hinweg zwischen dem Anteil der 14 gemeldeten Tage, deren Verlust unter den Schwellenwert fällt, und dem Anteil der 17 anderen Tage, deren Verlust dies tut, für eine Hypothese, die aus denselben gemeldeten Tagen mit d Merkmale gelernt wurde. Daten generiert von `pythondemos/membershipinferenceattack.py`

Der attack ist nur insoweit erfolgreich, als die gelernte Hypothese einen geringeren Verlust innerhalb des Trainingsdatensatzs als außerhalb davon verursacht, was der generalization gap misst. Unter einem Wahrscheinlichkeitsmodell dafür, wie die Datenpunkte und ihre Zugehörigkeit zum Trainingsdatensatz entstehen, ist der Erfolg eines solchen attacks an diese Lücke gebunden. Die Größe ist die gleiche Differenz wie oben, wobei die beiden Raten durch Wahrscheinlichkeiten unter diesem Modell ersetzt werden. Für einen Verlust, der durch eine Konstante B begrenzt ist, hat ein Angreifer, der mit einer im Verlust abnehmenden Wahrscheinlichkeit mit Ja antwortet, eine Differenz, die proportional zum erwarteten generalization gap wächst (Yeom et al., 2018, Thm. 2). Die Regularisierung kontrolliert diese Lücke: Für einen konvexen Verlust, der Lipschitz mit der Konstanten L ist, begrenzt das Hinzufügen des Strafterms $\alpha\|\mathbf{w}\|_2^2$ den erwarteten generalization gap durch $2L^2/(\alpha m)$ (Shalev-Shwartz and Ben-David,

2014, Cor. 13.6). Abb. 3 wiederholt den attack gegen dieselben Merkmale, die mit diesem Strafterm gelernt wurden, und das Bild unterscheidet die beiden Gruppen nicht mehr. Die Regularisierung dient daher zwei Zwecken: genauen Vorhersagen außerhalb des Trainingsdatensatzs (Hastie et al., 2009, Sect. 7.4) und einer Verteidigung gegen einen Membership-Inference-attack (Shokri et al., 2017, Sect. VIII). Für den Beobachter in Krems ist das der Unterschied zwischen einer veröffentlichten Hypothese, die nichts über ihn preisgibt, und einer, die die Tage preisgibt, an denen er in der Stadt war.

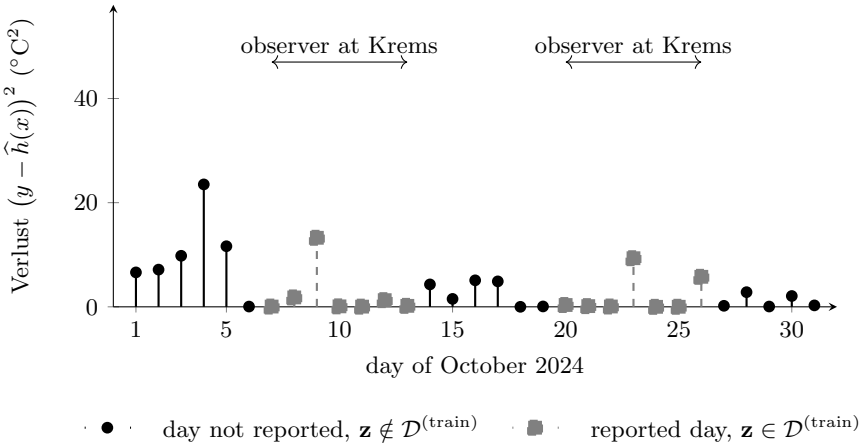

Abbildung 3: Derselbe attack wie in Abb. 1, auf denselben Achsen, gegen die Hypothese, die aus denselben 14 gemeldeten Tagen mit dem Strafterm $\alpha\|\mathbf{w}\|_2^2$ für $\alpha = 10$ gelernt wurde. Die gemeldeten Tage tragen nun gestrichelte Stiele, die von einem Quadrat gekrönt sind, von ungefähr gleicher Höhe wie die durchgezogenen Stiele, die von einer Scheibe gekrönt sind und die Tage markieren, die der Beobachter nicht gemeldet hat, sodass kein Schwellenwert sie hervorhebt. Ihr durchschnittlicher Verlust steigt von 0 auf 2.31 °C², während der durchschnittliche Verlust anderswo auf 4.71 °C² fällt: Der generalization gap schließt sich auf 2.40 °C² und der Vorteil des Angreifers sinkt von eins auf höchstens 0.37, der größte Teil davon ist verschwunden, aber nicht alles. Daten generiert von `pythondemos/membershipinferenceattack.py`

Querverweise

attack, privacy attack, personal data, Regularisierung, DP, Überanpassung, Trainingsfehler, Validierungsfehler, Trainingsdatensatz

Literatur

1. European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the

processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance). Official Journal of the European Union, L 119/1, May 4. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj.

2. Yeom et al. (2018). Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In: 2018 IEEE 31st Computer Security Foundations Symposium (CSF), pp. 268–282.

3. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.

4. Shokri et al. (2017). Membership Inference Attacks Against Machine Learning Models. In: 2017 IEEE Symp. Secur. Privacy, pp. 3–18.

5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.

6. Vassilev et al. (2024). Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, Gaithersburg, MD.