Vor dem erfolgreichen Machine Learning kommt die Datenjagd

Adarsh Kyadige, Sebior Data Scientist im KI-Team von Sophos nennt erste Schritte und Unwägbarkeiten beim Erstellen von Machine-Learning-Modellen

Künstliche Intelligenz bzw. Machine Learning hat in den letzten zehn Jahren einen enormen Aufschwung erlebt. Viele Branchen investieren jetzt massiv in Lösungen, die auf maschinellem Lernen basieren. Auch die Nachfrage nach qualifizierten Spezialisten ist sprunghaft gestiegen. Mehrere Universitäten weltweit bieten Abschlüsse mit dem Schwerpunkt Data Science oder Künstlicher Intelligenz an, und auch an deutschen Hochschulen gewinnen diese Inhalte deutlich an Bedeutung.

Während sich Universitäten dabei vor allem auf die mathematischen und theoretischen Konzepte konzentrieren, können die erforderlichen Fähigkeiten und Kenntnisse für das Training von Machine-Learning-Modellen bei Problemstellungen in der realen Welt ganz anders aussehen.

Verfügbarkeit der notwendigen Daten
In den meisten Fällen entscheidet die Verfügbarkeit von Daten darüber, ob maschinelles Lernen zur Lösung eines bestimmten Problems eingesetzt werden kann oder nicht. Vor dem Start eines neuen Projekts steht daher die Frage: Wird ein auf diesen Daten trainiertes Modell die meiste Zeit die richtigen Antworten liefern?

Diese Frage gilt unabhängig von dem Modell, der Bibliothek oder der Sprache, die für das ML-Experiment gewählt wird. Und es gibt weitere entscheidende Kriterien. Ein Modell ist immer nur so gut, wie die Daten, die ihm zugeführt werden. Wichtig ist daher zu klären:

- Sind ausreichend Daten vorhanden, um ein gutes Modell zu trainieren? Sofern es das Hardware-Budget nicht überschreitet, ist es fast immer richtig, mehr Daten zu verwenden.
- Sind die Prognosen bei einem überwachten Lernprozess belastbar? Wird das Modell mit den richtigen Informationen gefüttert?
- Sind diese Daten eine genaue Darstellung der realen Verteilung? Sind genügend Variationen in den Proben, um den Problembereich abzudecken?
- Besteht konstanter Zugang zu einem ständigen Strom neuer Daten, mit denen das Modell aktualisiert und auf dem neuesten Stand gehalten werden kann?

Zusammenstellen der Daten
Die notwendigen Daten für die Erstellung eines Datensatzes für eine ML-Lösung befinden sich häufig verteilt auf mehrere Quellen. Verschiedene Teile einer Stichprobe werden über verschiedene Produkte gesammelt und von diversen Teams auf unterschiedlichen Plattformen verwaltet. Daher besteht der nächste Schritt im Prozess oft darin, all diese Daten in einem einzigen Format zusammenzufassen und so zu speichern, dass sie leicht zugänglich sind.

Weitere Herausforderungen und ein Fluch
Wenn die Daten gesammelt und aggregiert sind, würde man meinen, dass es nun losgehen könnte mit dem fabelhaften neuen ML-Algorithmus. Doch noch sind weitere Schritte notwendig, denn es werden unweigerlich noch einige Herausforderungen zu bewältigen sein:

Fehlende Daten: Manchmal sind vielleicht nicht für alle Beobachtungen gültige Werte verfügbar. Daten könnten während der Sammlung, Speicherung oder Übertragung beschädigt worden sein, und es gilt, diese fehlenden Datenpunkte zu finden und sie ggfs. aus dem Datensatz zu löschen.

Doppelte Daten: Auch wenn dies im Hinblick auf die Modell-Performance kein besonders alarmierendes Problem darstellt, sollten doppelte Daten aus dem Datenspeicher entfernt werden, um den Modelltrainingsprozess effizienter zu gestalten und möglicherweise eine Überanpassung zu vermeiden.

Verschiedene Normalisierungsschemata: Geringe Unterschiede in der Art und Weise, wie die Daten verarbeitet und gespeichert werden, können beim Training eines Modells große Kopfschmerzen verursachen. Beispielsweise können verschiedene Produkte dasselbe Freitextfeld auf unterschiedliche Längen beschneiden oder Daten unterschiedlich anonymisieren, was zu Inkonsistenzen in den Daten führt. Wenn eine dieser Quellen überwiegend Malware und eine andere Quelle gutartige Muster enthält, kann das ML-Modell lernen, sie z.B. anhand der Beschnittlänge zu identifizieren.

Freitextfelddaten: Dies verdient eigentlich eine Kategorie für sich allein, weil es so schwierig sein kann, damit umzugehen. Freitextfelder sind der Fluch des Daten-Ingenieurs, da er sich mit Tippfehlern, Umgangssprache, Beinahe-Duplikaten, Variationen in der Groß- und Kleinschreibung, Leerzeichen, Interpunktion und einer ganzen Reihe anderer Inkonsistenzen auseinandersetzen muss.

Stetige Aktualisierung

Die Datendrift schließlich ist ein wichtiges Problem, das beim Entwurf eines ML-Systems angegangen werden muss. Sobald ein Modell trainiert ist, wird es im Laufe der Zeit immer ungenauer, da sich die Verteilung der neu eingehenden Daten ändert. Daher sollte eine regelmäßige Aktualisierung des Modells festgelegt werden, um sicherzustellen, dass die Leistung weiterhin innerhalb der erwarteten Grenzen liegt.
Im Sicherheitsbereich sehen wir zum Beispiel eine große Volatilität, da Bedrohungsakteure ihre Exploits und ihr Verhalten im Laufe der Zeit ändern und Schwachstellen entdeckt und behoben werden.

Dies war eine kurze Zusammenfassung der typischen Schritte, die unternommen werden müssen, um Daten für eine ML-Lösung auszuwählen, zu sammeln und zu bereinigen. Sind diese alle erfolgt, steht vermutlich ein sauberer Datensatz zur Verfügung.
Das Experiment kann beginnen.

Pressekontakt:

Sophos
Jörg Schindler, PR-Manager Central & Eastern Europe
joerg.schindler@sophos.com, +49-721-25516-263

TC Communications
Arno Lücht, +49-8081-954619
Thilo Christ, +49-8081-954617
Ulrike Masztalerz, +49-30-55248198
Ariane Wendt +49-172-4536839
sophos@tc-communications.de