“Why Should I Trust You?” Explaining the Predictions of Any Classifier [1]

Streszczenie: Dariusz Max Adamski

1. Motywacja

Powszechne zastosowanie systemów uczenia maszynowego uwidoczniło problemy związane z ich często niską wiarygodnością. Aby doprecyzować kwestię wiarygodności modeli, autorzy wymieniają dwie definicje.

W pierwszej definicji, predykcja modelu jest wiarygodna, jeśli użytkownik modelu ufa danej predykcji wystarczająco, żeby podjąć na jej podstawie jakąś decyzję. Wiarygodność predykcji jest szczególnie istotna, gdy decyzje na podstawie pojedynczych predykcji muszą mieć mocne uzasadnienie (medycyna, bezpieczeństwo).

W drugiej definicji, model jest wiarygodny, jeśli użytkownik ufa, że predykcje modelu na danych spoza zbioru danych będą sensowne. Autorzy zaznaczają, że obliczenie wartości wybranych miar jakości na walidacyjnym zbiorze danych nie jest wystarczające, ponieważ rzeczywiste dane często znacznie się różnią od tych w treningowym i walidacyjnym zbiorze danych.

2. Metody

Jako rozwiązanie problemu wiarygodności predykcji, autorzy proponują metodę LIME, która wyjaśnia pojedyncze predykcje dowolnych modeli klasyfikacji i regresji, a jako rozwiązanie problemu wiarygodności modelu, proponują metodę SP-LIME, która wybiera reprezentatywne przykłady ze zbioru danych i dostarcza ich wyjaśnienia.

Autorzy definiują cztery pożądane cechy metod wyjaśniania. Po pierwsze, wyjaśnienia dostarczone przez metodę muszą być interpretowalne, czyli zrozumiałe dla użytkownika. Po drugie, wyjaśnienia powinny być lokalnie wierne, czyli odzwierciedlać predykcje modelu w otoczeniu konkretnego przykładu. Po trzecie, metoda wyjaśniania powinna traktować model jako czarną skrzynkę, czyli działać niezależnie od użytego algorytmu uczącego. Ostatnią pożądaną cechą jest możliwość globalnej oceny wiarygodności modelu.

Wyjaśnienia zaproponowanej przez autorów metody LIME (Local Interpretable Model-Agnostic Explanations), posiadają trzy z czterech wyżej wymienionych cech: są interpretowalne, lokalnie wierne i niezależne od modelu. Interpretowalność i lokalna wierność wyjaśnień LIME jest zapewniona przez zdefiniowanie funkcji celu jako sumy miary niewierności wyjaśnianego modelu do wyjaśniającego modelu w otoczeniu wybranego przykładu, oraz miary złożoności wyjaśniającego modelu. Niezależność od wyjaśnianego modelu jest zapewniona przez brak założeń o wyjaśnianym modelu w mierze niewierności. Wyjaśnienia metody SP-LIME (Submodular Pick - LIME) dodatkowo umożliwiają globalną

ocenę wiarygodności modelu przez dostarczenie wyjaśnień dla reprezentatywnego zbioru przykładów.

Aby zaprezentować działanie metody LIME, autorzy tworzą wyjaśnienia dla modelu SVM w zadaniu klasyfikacji tekstu. Działanie LIME zostało też pokazane na przykładzie wyjaśnienia klasyfikacji obrazów przez pre-trenowaną sieć neuronową Inception.

3. Eksperymenty

W celu sprawdzenia efektywności wprowadzonych metod, autorzy przeprowadzili eksperymenty z udziałem użytkowników oraz eksperymenty z symulowanymi użytkownikami.

W eksperymentach z symulowanymi użytkownikami sprawdzane były wierność modeli wyjaśniających LIME, wiarygodność wyjaśnień oraz globalna wiarygodność modelu. Eksperymenty były przeprowadzane na różnych modelach (decision trees, logistic regression, random forest, itp.) w zadaniu klasyfikacji sentymentu recenzji książek i filmów. Metodę LIME porównano z metodą Parzen, czyli metodą wyjaśniania globalnym przybliżeniem modelu, opartym na oknach Parzena. W eksperymencie badającym wiarygodność wyjaśnień, metoda LIME uzyskała istotnie lepsze wyniki od metody Parzen.

Eksperymenty z udziałem prawdziwych użytkowników polegały na sprawdzeniu, czy metoda SP-LIME pozwala użytkownikom wybrać lepszy z dwóch modeli, czy uzyskane wyjaśnienia wspomagają dobór cech, oraz czy wyjaśnienia pomagają zidentyfikować niechciane cechy w klasyfikatorach. W pierwszym eksperymencie, metoda SP-LIME pozwoliła użytkownikom wybrać lepszy z dwóch modeli ze średnią dokładnością 89%. W drugim eksperymencie, LIME pozwoliła użytkownikom niebędącym ekspertami zwiększyć dokładność klasyfikatora przez dobór cech. W trzecim eksperymencie wytrenowano intencjonalnie słaby klasyfikator obrazków. Wyjaśnienia LIME zgodnie z oczekiwaniami obniżyły zaufanie do słabego klasyfikatora.

4. Podsumowanie

Wprowadzone przez autorów metody posiadają pożądane według autorów cechy wyjaśnialności: interpretowalność, lokalna wierność, uniwersalność i możliwość globalnej oceny modelu. W eksperymentach, autorzy pokazali, że LIME i SP-LIME pozwalają użytkownikom dokonywać selekcji modeli, selekcji cech, a wyjaśnienia predykcji mogą obniżyć zaufanie do słabych modeli.

W perspektywie czasu, metoda LIME okazała się użytecznym narzędziem, które aktualnie jest używane do tworzenia bardziej wiarygodnych modeli uczenia maszynowego. Metoda LIME stała się bardzo popularna i ma implementacje w wielu językach programowania, dzięki czemu kontrola wiarygodności modeli uczenia maszynowego jest bardziej powszechna.

Bibliografia

[1] Marco Tulio Ribeiro, Sameer Singh i Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. W: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD’16. San Francisco, California, USA: Association for Computing Machinery, 2016, s. 1135–1144. isbn: 9781450342322. doi: 10.1145/2939672.2939778. https://arxiv.org/pdf/1602.04938.pdf.