Gniewosz **Leliwa**

SZTUCZNA INTELIGENCJA

O czym myśli, gdy nikt nie patrzy?

Helion

Redaktor prowadzący: Małgorzata Kulik

Projekt okładki: Studio Gravite/Olsztyn
Obarek, Pokoński, Pazdrijowski, Zaprucki
Grafika na okładce została wykorzystana za zgodą AdobeStock.com.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: helion@helion.pl
WWW: helion.pl (księgarnia internetowa, katalog książek)

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
helion.pl/user/opinie/sztuit
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

ISBN: 978-83-289-2310-2

Printed in Poland.

SPIS TREŚCI

KLĄTWA CZARNEJ SKRZYNKI

Dokonało się — oto zerwaliśmy owoc z drzewa poznania sieci neuronowych. Posiedliśmy wiedzę, która pozwoli nam wreszcie odpowiedzieć na pytanie z początku książki. To, które zadał młody Symbolista, a o którym większość Czytelników zdążyła już zapewne zapomnieć: dlaczego nikt nie potrafił wyjaśnić działań robota Koneksjonistów, który twierdził, że na zdjęciu jest kot, pomimo że go tam nie było?

Teraz już wiemy, że robot Koneksjonistów, czyli wyuczona sieć neuronowa, to nic innego jak zbiór bardzo wielu połączonych ze sobą neuronów, umieszczonych na kolejnych warstwach sieci, którym dobrano odpowiednie wagi w procesie jej uczenia. Ale czy umielibyśmy powiedzieć, dlaczego na wyjściu dostajemy taką, a nie inną wartość?

Moglibyśmy spróbować prześledzić wszystkie połączenia między neuronami i ich wagi, żeby sprawdzić, które sygnały wejściowe są wzmacniane, a które wygaszane. I być może dla bardzo prostych przypadków w drodze dedukcji udałoby się ustalić, które z nich miały największy wpływ na wyjście sieci. Ale im głębsza sieć, tym bardziej rozmywają się te wszelkie zależności i tym trudniej o jakiekolwiek sensowne wnioski. Spróbuję to wyjaśnić na przykładzie.

Wyobraźmy sobie, że działanie sieci neuronowej jest jak układanie puzzli:

1. Warstwa wejściowa reprezentuje puzzle wysypane z pudełka na podłogę.
2. W pierwszej warstwie ukrytej odwracamy wszystkie elementy obrazkiem do góry.
3. W drugiej szukamy brzegów i układamy „ramkę".
4. W trzeciej układamy kształty na zasadzie kontrastu, np. dachy na tle nieba.

5. W czwartej łączymy te kształty w obiekty, np. ściany, okna i dachy w budynki.
6. W piątej łączymy ze sobą całe obiekty, np. budynki w nadmorskie miasteczko.
7. W szóstej wypełniamy to, co pozostało niewypełnione — niebo, wodę czy trawę.
8. W warstwie wyjściowej mamy już kompletny obraz i możemy w końcu powiedzieć, co się na nim znajduje.

Nie, wcześniej nie mogliśmy, bo pudełko z obrazkiem zjadł nam pies. Proszę się nie czepiać. I tak — zdaję sobie sprawę, że istnieją inne strategie układania puzzli. Wiem też, że wcale nie musieliśmy układać całości, żeby móc powiedzieć, co znajduje się na obrazku. Nie psujmy sobie jednak tak pięknej analogii jakimiś błahostkami.

Każda kolejna warstwa sieci to kolejny poziom abstrakcji — pojedyncze elementy, kształty, obiekty, grupy obiektów. Problem polega na tym, że kolejne warstwy ukryte rzeczywiście mogą reprezentować te nasze wyobrażenia, ale wcale nie muszą. Mogą reprezentować coś zupełnie innego. Coś, o czym nigdy byśmy nie pomyśleli jako przedstawiciele *homo sapiens* — gatunku, którego mózg, wbrew pozorom, wcale nie służy do myślenia. Tylko czy to aby na pewno jest problem?

Nikt przy zdrowych zmysłach nie powinien twierdzić, że nasz sposób postrzegania rzeczywistości jest jedynym możliwym albo obiektywnie lepszym od innych. Jeśli nie będziemy się upierać, że krzesło tym się różni od stołu, że krzesło ma oparcie, a stół nie, to nigdy nie natrafimy na problem w postaci taboretu. I bardzo dobrze. Wystarczy, że na co dzień bohatersko walczymy z całą masą problemów, które sami stworzyliśmy.

Dlatego fakt, że nie musimy tłumaczyć maszynie, jak wygląda kot ani czym różni się od psa, uważam za ogromną zaletę. Sieć sama nauczy się tego z pokazywanych jej przykładów. Nie potrzebujemy dyplomu z anatomii zwierząt domowych. Nie potrzebujemy głębokiej wiedzy domenowej. Ba, w wielu dziedzinach nawet jej nie mamy. Albo, co gorsza, nasze przyzwyczajenia lub wyobrażenia na dany temat są wręcz szkodliwe.

W latach 2004 – 2012 amerykańska stacja telewizyjna Fox emitowała bardzo popularny, również w Polsce, serial o lekarzach — *Dr House*. Większość odcinków do znudzenia eksploatowała następujący schemat — pacjent choruje na pewną chorobę A, ale symptomy i pierwsze wyniki badań wskazują na inną, bardziej prawdopodobną — przynajmniej zdaniem większości lekarzy — chorobę B. Pacjent zaczyna być leczony zgodnie z tą drugą diagnozą i wkrótce jego stan się pogarsza.

I tu wchodzi on — tytułowy doktor House. Cały na biało, mimo że akurat jako jedyny lekarz w serialu nie nosi fartucha. Łączy pozornie nieistotne fakty — bardzo rzadki objaw choroby lub skutek uboczny działania leku ze specyficzną wiedzą o pacjencie, zazwyczaj pozyskaną z… włamania do jego domu — i stawia właściwą diagnozę. Pacjent zaczyna być leczony na A, dochodzi do siebie i odjeżdża w stronę zachodzącego słońca. Napisy końcowe.

Ale po co ja o tym wszystkim piszę? Bo wszyscy kłamią? Nie, nie tym razem. Otóż nasza sieć neuronowa, właśnie dlatego, że nie obarczamy jej naszymi przypuszczeniami czy uprzedzeniami, może być takim doktorem House'em na sterydach. Może dostrzec rzadkie zależności, których młody lekarz nie zauważy, a lekarz z wieloletnim doświadczeniem zignoruje jako mało prawdopodobne. Albo analizując setki tysięcy przypadków, znajdzie zależności, których żaden lekarz nigdy wcześniej zwyczajnie nie mógł dostrzec ze względu na skalę. Do najciekawszych zastosowań sztucznej inteligencji jeszcze wrócimy, ale teraz porozmawiajmy w końcu o klątwie czarnej skrzynki, bo niestety każdy kij ma dwa końce.

***

Czarna skrzynka to termin określający system lub urządzenie o znanej funkcjonalności, ale nieznanym mechanizmie działania. Znamy wejście, znamy wyjście, ale nie umiemy powiedzieć, co dokładnie dzieje się w środku. Wyuczona sieć neuronowa jest właśnie taką czarną skrzynką.

Zgoda — mamy warstwy i neurony, wiemy, ile ich jest, możemy nawet podejrzeć wagi na poszczególnych połączeniach. Ale nie mamy pojęcia, co reprezentują kolejne warstwy ukryte sieci. Nie umiemy powiedzieć, dlaczego robot Koneksjonistów stwierdził, że na zdjęciu jest kot, pomimo że go tam nie było. Możemy

zgadywać, że model pomylił kształt kota z kształtem poduszki, ale tak po prawdzie nie wiemy nawet, czy on w ogóle szukał jakichkolwiek kształtów.

To dość paradoksalne, bo wiemy, że sztuczne neurony i tworzone z nich sieci są wzorowane na ich biologicznych odpowiednikach. Być może stąd nasza przemożna chęć do antropomorfizacji, do doszukiwania się w nich naszych ludzkich zachowań i ludzkiego sposobu postrzegania rzeczywistości. To błąd, który może nas wpuścić w maliny. Lepiej jest wiedzieć, że czegoś nie wiemy, i zaakceptować ten fakt, niż wyciągać daleko idące wnioski na podstawie naszych wątpliwych przypuszczeń.

Zastanówmy się zatem — co to dla nas oznacza? Czy to rzeczywiście problem, że nie wiemy, co dzieje się w środku czarnej skrzynki? Przy tak postawionym pytaniu na usta ciśnie się najbardziej naukowo-inżynierska odpowiedź, jaką na przestrzeni wieków podawały najtęższe umysły tego świata, czyli: to zależy.

A od czego, jeśli można wiedzieć? Przede wszystkim od zastosowania. Nikt nie będzie rwał sobie włosów z głowy, jeśli nie dowiemy się, dlaczego model pomylił kota z poduszką. Albo dlaczego błędnie ocenił negatywną recenzję jako pozytywną. Zwłaszcza że w przypadku takich zastosowań jak analiza sentymentu istotna jest statystka, a nie pojedyncze wpadki modelu. Zatrzymajmy się tu na moment i porozmawiajmy o tym, w jaki sposób testuje się modele uczenia maszynowego.

***

Z poprzednich rozdziałów wiemy, że sieci neuronowe uczy się, „pokazując" im kolejne przykłady ze zbioru uczącego. Testuje się bardzo podobnie — „pokazując" analogiczne przykłady ze zbioru testowego. Oba zbiory powinny być rozłączne, co oznacza, że model testuje się wyłącznie na przykładach, które nie były wcześniej wykorzystywane do uczenia sieci.

Zazwyczaj przygotowuje się jeden duży zbiór danych, który następnie dzieli się na zbiór uczący i testowy w ustalonych proporcjach, np. 70% przypadków trafia do zbioru uczącego, a 30% do testowego. Na podstawie przypadków testowych sprawdza się poprawność predykcji modelu, czyli zgodność wartości wyjściowej z wartością oczekiwaną.

W przypadku problemów klasyfikacyjnych wyróżniamy 4 typy wyników, które możemy otrzymać, oceniając pojedynczy przypadek testowy. Znowu muszę poprosić o chwilę uwagi i skupienia. Za moment wprowadzimy kilka nowych oznaczeń, wzorów i dziwacznych określeń, ale Szanowny Czytelnik przekona się, że to wszystko ma sens.

Dlaczego aż 4 typy wyników? Dlatego, że mamy ocenę modelu, którą sprawdzamy, i ocenę człowieka, którą traktujemy jako poprawną. Każdy przypadek może należeć (wynik dodatni) lub nie należeć (wynik ujemy) do danej kategorii. I każdą z tych dwóch możliwości przyrównujemy do oceny człowieka, będącej jedynym źródłem prawdy. Jeśli ocena modelu zgadza się z oceną człowieka, to mamy wynik prawdziwy, a jeśli się nie zgadza — fałszywy.

Stąd oceniając poprawność klasyfikacji, możemy otrzymać 1 z 4 wyników:

- wynik prawdziwie dodatni (*TP* od angielskiego *true positive*), kiedy przypadek testowy należy do danej kategorii i model prawidłowo go do niej zaklasyfikował;
- wynik prawdziwie ujemy (*TN* od angielskiego *true negative*), kiedy przypadek testowy NIE należy do danej kategorii i model prawidłowo go do niej NIE zaklasyfikował;
- wynik fałszywie pozytywny (*FP* od angielskiego *false positive*), nazywany również błędem I typu, kiedy przypadek testowy NIE należy do danej kategorii, ale model go do niej błędnie zaklasyfikował;
- wynik fałszywie negatywny (*FN* od angielskiego *false negative*), nazywany również błędem II typu, kiedy przypadek testowy należy do danej kategorii, ale model błędnie go do niej NIE zaklasyfikował.

Poniższa tabelka przedstawia wszystkie możliwe wyniki dla prostego modelu oceniającego, czy na zdjęciu jest kot, czy też go tam nie ma.

Tabela 1. Wszystkie możliwe do uzyskania wyniki oceny modelu rozpoznającego na zdjęciu kota lub jego brak

OCENA CZŁOWIEKA	OCENA MODELU	WYNIK
kot	*kot*	*TP — prawdziwie dodatni*
nie kot	*nie kot*	*TN — prawdziwie ujemny*
nie kot	*kot*	*FP — fałszywie dodatni*
kot	*nie kot*	*FN — fałszywie ujemny*

Jeśli teraz dla całego zbioru testowego podliczymy poszczególne typy otrzymywanych wyników, to będziemy mogli policzyć dwie nowe i piekielnie użyteczne metryki — precyzję (od angielskiego *precision*) oraz kompletność (od angielskiego *recall*). Ze względu na ich charakter lubimy przedstawiać je procentowo, ale może po kolei…

W prostych żołnierskich słowach: jeśli mamy koszyk i chcemy nazbierać do niego jabłek, to kompletność określa, ile spośród wszystkich jabłek udało nam się zebrać, a precyzja pozwala określić, ile gruszek i śliwek przez przypadek trafiło do naszego koszyka z jabłkami.

Kompletność mówi nam, ile ze wszystkich przypadków, które powinny zostać zaklasyfikowane do danej kategorii, rzeczywiście do niej trafiło. Jeśli do zebrania było 10 jabłek, a my znaleźliśmy tylko 8 z nich, to kompletność wynosić będzie 80%.

Precyzja natomiast określa, ile przypadków, które zaklasyfikowaliśmy do danej kategorii, słusznie do niej trafiło. Jeśli do koszyka trafiło 10 owoców, a tylko 8 z nich to jabłka, to precyzja także wynosić będzie 80%.

Do wyznaczenia precyzji i kompletności wystarczy nam suma poszczególnych typów wyników w zbiorze testowym. Niech *TP* oznacza liczbę wyników prawdziwie dodatnich, *FP* — liczbę wyników fałszywie dodatnich, a *FN* — liczbę wyników fałszywie ujemnych. Wtedy możemy zapisać następujące wzory:

- $precyzja = \frac{TP}{TP+FP}$
- $kompletność = \frac{TP}{TP+FN}$

A skoro tak dobrze nam idzie, to zapiszmy od razu wzór na średnią harmoniczną F_1, łączącą obie te metryki:

- $F_1 = 2 \times \frac{precyzja \times kompletność}{precyzja + kompletność}$

Po co nam ona? — zapytacie. I słusznie, bo naprawdę dobry inżynier zamiast dodawać kolejne byty, powinien odejmować je tak długo, aż nie pozostanie już nic więcej do odjęcia. Ta średnia odegra jednak dość ważną rolę, bo posiada tę interesującą cechę, że przyjmuje najwyższe wartości wtedy, kiedy obie metryki są sobie równe, a w przypadku rosnących dysproporcji gwałtownie spada.

Policzmy:

- 50% precyzji i 50% kompletności daje $F_1 = 50\%$,
- 70% precyzji i 30% kompletności daje $F_1 = 42\%$,
- 90% precyzji i 10% kompletności daje $F_1 = 18\%$.

I z tej przyczyny to właśnie pod średnią harmoniczną precyzji i kompletności najczęściej optymalizuje się modele uczenia maszynowego. Stosuje się do tego trzeci zbiór danych, tzw. zbiór walidacyjny, który znów trzeba „wyszarpać" ze zbioru uczącego i testowego, choćby w jednej z popularnych proporcji — 80/10/10, 70/15/15 lub 60/20/20, gdzie oczywiście dominującym składnikiem jest zbiór uczący.

***

Nie wystarczy zatem, że nauczymy sieć, jak rozwiązywać dany problem. Musimy ją jeszcze porządnie przetestować, żeby wiedzieć, na ile wiarygodne są jej predykcje. Dopiero wtedy będziemy w stanie ocenić, czy i w jaki sposób możemy wykorzystać nasze rozwiązanie, godząc się przy tym, że nie będziemy wiedzieli, co dzieje się w jego wnętrzu i w jaki dokładnie sposób podejmowana jest decyzja.

Jeśli do analizy sentymentu zastosujemy model uzyskujący w testach 90% precyzji i 70% kompletności, to możemy założyć, że trend prezentowany przez ten model będzie wiarygodny, i możemy kierować się nim przy podejmowaniu naszych decyzji biznesowych. Jeśli stosunek recenzji pozytywnych do negatywnych rośnie w czasie, to nie ma powodu do zmartwień. Jeśli spada — warto zainteresować się tematem.

A jeśli nagle zalewa nas fala negatywnych recenzji, to trzeba czym prędzej bić na alarm, bo najprawdopodobniej coś poszło nie tak i wymagana jest natychmiastowa reakcja.

Jak można się domyślać, istnieją również zastosowania, dla których wyjaśnialność modelu jest istotna lub wręcz kluczowa. Jeśli działamy w oparciu o trend lub statystkę, to — jak w powyższym przypadku — możemy po prostu zamknąć oczy i zaufać modelowi. Jeśli jednak nasze działania zależą od wyniku pojedynczej predykcji, to wolelibyśmy rozumieć, skąd taka, a nie inna decyzja, a nie polegać wyłącznie na ogólnej, choćby nie wiem jak wysokiej i dokładnie zmierzonej skuteczności modelu. A im większa waga tej decyzji i podejmowanego w oparciu o nią działania, tym większe zapotrzebowanie na wyjaśnialność modelu. W skrajnych przypadkach może przecież chodzić nawet o ludzkie życie.

W naturalny sposób przychodzą tu na myśl zastosowania sztucznej inteligencji w medycynie. Jeśli nasz model zdiagnozował u pacjenta nowotwór, to chcemy zrozumieć, skąd taka diagnoza, zanim zaczniemy stosować agresywne metody leczenia lub interwencję chirurgiczną. Oczywiście w przypadku modeli diagnostycznych mamy lekarza, który może poszukać wyjaśnienia poza modelem. A co, jeśli nie ma na to czasu i decyzję trzeba podjąć natychmiast, wyłącznie w oparciu o predykcję modelu?

Że niby trudno wyobrazić sobie taką sytuację? A pojazdy autonomiczne? Samochody samosterujące? Prace nad nimi trwają już od wielu lat, a od co najmniej kilku są z powodzeniem testowane w Stanach Zjednoczonych, w Kalifornii i Arizonie, oraz w Chinach, na ulicach liczącego ponad 13 milionów mieszkańców miasta Shenzhen. Ich decyzje muszą być podejmowane autonomicznie, w czasie rzeczywistym, w oparciu o sygnały spływające z różnego rodzaju czujników. Taki pojazd musi w ułamku sekundy ocenić czy to, co nagle pojawiło się na jezdni, to pieszy, czy może targany wiatrem worek na śmieci. Ocenić i zareagować, nie stwarzając przy tym niebezpieczeństwa dla pozostałych uczestników ruchu drogowego.

Tyle że to znowu przypadek skrajny, bo — ze względu na wymagany czas reakcji — ewentualna wyjaśnialność modelu raczej nie zapobiegłaby nieszczęściu, a jedynie mogłaby pomóc w ustaleniu, dlaczego do nieszczęścia w ogóle doszło. No więc kiedy cała ta wyjaśnialność ma największe znaczenie? Wtedy, kiedy okienko

czasowe na podjęcie decyzji jest na tyle duże, że pozwala na uwzględnienie wyjaśnień predykcji, ale na tyle małe, że nie pozwala na przeprowadzenie niezależnej analizy poza modelem. Fani filmów akcji i dreszczowców bez pudła wskażą tu przeróżne zastosowania militarne, gdzie przesadna zwłoka w podejmowaniu decyzji może prowadzić do bardzo poważnych konsekwencji.

Wyjaśnialność modelu ma również znaczenie wtedy, gdy decyzję musimy podjąć natychmiast, ale później musimy się z niej jeszcze wytłumaczyć. Tu z przyjemnością podzielę się przykładami z mojej pracy zawodowej, z firmy Samurai Labs, której jestem dumnym współzałożycielem i w której — jako dyrektor ds. technologii — odpowiadam między innymi za tematy związane ze sztuczną inteligencją. Jednym z rozwiązań oferowanych przez Samurai Labs jest proaktywna ochrona społeczności internetowych przed całą masą niebezpiecznych zjawisk i zachowań, szeroko nazywanych cyberprzemocą.

Chyba każdy z nas korzystał kiedyś z jakiegoś komunikatora internetowego, czatu, forum albo wymieniał opinie w komentarzach czy to pod filmem na YouTubie, czy też w mediach społecznościowych. Gdy pojawia się gagatek, który w rozmowie obraża lub nęka innych i nie szanuje zasad panujących w danej społeczności, to szybka reakcja pozwala uchronić potencjalne ofiary i zapobiec eskalacji przemocy. Dlatego zadaniem naszej sztucznej inteligencji, pełniącej wtedy rolę autonomicznego moderatora, jest reagować natychmiast, proaktywnie, zanim komukolwiek stanie się krzywda.

Ale ponieważ funkcjonujemy również w otwartych społecznościach internetowych, gdzie wolność słowa jest cenioną wartością, to ów gagatek ma prawo wiedzieć, za co dokładnie został czasowo wyciszony lub zbanowany albo dlaczego jego wypowiedź została zablokowana lub usunięta. Jeśli jesteśmy w stanie podjąć natychmiastową decyzję i skutecznie wytłumaczyć nasze przesłanki, to takie podejście ma dodatkowe walory edukacyjne, deeskalacyjne i pozwala na utrzymanie zdrowych standardów komunikacji — bez cyberprzemocy, ale i bez cenzury.

Innym bardzo ważnym i niezwykle trudnym zagadnieniem, którym zajmujemy się w Samurai Labs, jest pomoc osobom w kryzysie presuicydalnym i suicydalnym, czyli osobom rozważającym lub planującym odebranie sobie życia. Tylko w 2023 roku nasza sztuczna inteligencja pozwoliła na udzielenie wsparcia ponad 25 tysiącom

takich osób na całym świecie. Wystarczy, że osoba w kryzysie da temu wyraz, opisując, co czuje, myśli lub przeżywa, na platformie internetowej, którą monitorujemy. Możemy wtedy wykryć zagrożenie i zareagować, wysyłając specjalną interwencję, wybrane przez ekspertów materiały samopomocowe oraz namiary na sprawdzone organizacje, które oferują profesjonalną, bezpłatną i anonimową pomoc.

I tu również, oprócz szybkiej reakcji, znaczenie ma wyjaśnialność predykcji, która pozwala lepiej ocenić stan osoby w kryzysie, a przez to skuteczniej dobrać metody dotarcia do niej i formy oferowanej pomocy. Istotnym predyktorem, czyli czynnikiem objaśniającym, jest ocena, czy mamy do czynienia z ideacją, zamiarem, czy planem. Osoba w kryzysie może dzielić się swoimi myślami samobójczymi, może pisać o zamiarze, gotowości i podjętej decyzji, a może przedstawiać szczegółowy opis uwzględniający czas, miejsce i sposób przeprowadzenia planowanej próby samobójczej. Rozróżnienie tych sytuacji jest niezwykle ważne. Do innych predyktorów możemy zaliczyć opis traumatycznych przeżyć, poprzednich prób samobójczych czy też akt pożegnania się z rodziną, przyjaciółmi lub innymi członkami danej społeczności.

Chyba wypada mi w tym momencie przeprosić Szanownego Czytelnika za tę niespodziewaną bombę. Zdaję sobie sprawę, że atmosfera zrobiła się już tak ciężka i gęsta, że można by ją kroić nożem, ale zależało mi na tym, żeby pokazać, że sztuczna inteligencja to nie tylko lepiej dobrane reklamy, ale także ratowanie ludzkiego życia. Tak na serio, czasem w ostatniej chwili. I tu nie tylko szybkość reakcji, ale również wyjaśnialność procesu podejmowania decyzji może mieć bardzo istotne znaczenie.

***

Chcemy mieć wyjaśnialne modele, bez dwóch zdań. Ale sieci neuronowe nie bardzo chcą z nami w tym aspekcie współpracować. Mimo to próbujemy, a tzw. wyjaśnialna sztuczna inteligencja (XAI, od angielskiego *eXplainable AI*) jest istotnym kierunkiem prac badawczych wielu firm i jednostek naukowych. Dlatego też sytuacja jest zła, ale nie beznadziejna.

Z pomocą przychodzą nam narzędzia, które wspierają nie tyle wyjaśnialność, co raczej interpretowalność modeli uczenia maszynowego. Jedne pozwalają przypisywać poszczególnym predykcjom tzw. stopień pewności (ang. *confidence score*), wskazujący, jak bardzo konkretne cechy wpływają na ostateczny wynik. Inne pomagają wykrywać negatywne cechy modeli, takie jak stronniczość (ang. *bias*) czy dryf danych (ang. *data drift*).

Stronniczość polega na tym, że wyuczony model wykazuje pewne systematyczne uprzedzenia w stosunku do określonej części analizowanych przypadków. Najczęściej dzieje się tak w przypadku źle dobranych lub źle zbalansowanych zbiorów uczących. Przykład: od wielu lat w amerykańskich szpitalach funkcjonuje system sztucznej inteligencji przewidujący, którzy pacjenci mogą wymagać dodatkowej opieki medycznej.

W 2019 roku wykryto, że algorytm wyraźnie dyskryminował pacjentów ciemnoskórych. Działo się tak dlatego, że model bazował na założeniu, że zapotrzebowanie na opiekę medyczną jest bezpośrednio skorelowane z indywidualnymi wydatkami na ochronę zdrowia. W związku z tym do uczenia modelu wykorzystano wcześniejsze wydatki pacjentów, a te okazały się wyraźnie niższe w przypadku osób ciemnoskórych. Podobną stronniczością wykazywały się algorytmy wspomagające rekrutację w dużych przedsiębiorstwach (dyskryminujące kobiety) lub oceniające zdolność kredytową potencjalnych klientów banków (dyskryminujące osoby ciemnoskóre).

Dryf danych jest związany z naturalną zmiennością świata, w którym przyszło nam żyć. To, że jakiś model wyuczony na danych z 2023 roku działa poprawnie dla danych z 2023 roku, nie oznacza, że będzie działał poprawnie dla analogicznych danych z 2024 lub 2025 roku. Świat nieustannie się zmienia i nasz model może osiągać gorsze wyniki dla nowych danych. Przykład: modele do rozpoznawania twarzy działały bardzo dobrze do 2020 roku, ale w związku z pandemią COVID-19 i obowiązkiem noszenia maseczek nagle przestały. Dlaczego? Bo w danych uczących nie było zdjęć twarzy w maseczkach, które od 2020 roku stanowiły absolutną większość danych wejściowych rejestrowanych w miejscach publicznych. Oczywiście taka „degradacja" modeli wcale nie musi być gwałtowna i skokowa. Może być powolna i stopniowa, np. wskutek długotrwałych zmian preferencji zakupowych konsumentów.

Warto dodać, że bardzo często metody stosowane w celu poprawy wyjaśnialności lub interpretowalności modeli uczenia maszynowego jednocześnie obniżają skuteczność samych modeli. W skrócie: im lepsza wyjaśnialność, tym niższa skuteczność. To trochę jak z zasadą nieoznaczoności Heisenberga, mówiącą, że istnieją pary wielkości, których nie da się zmierzyć z dowolną dokładnością. Im dokładniej znamy położenie cząstki, tym mniej wiemy o jej pędzie. A dla tych, którzy reagują alergicznie na wszelkie odniesienia do mechaniki kwantowej, mamy najlepsze możliwe podsumowanie tego zjawiska, zaserwowane nam przez brytyjskiego pisarza Alana Alexandra Milne'a w *Chatce Puchatka*: „Im bardziej Puchatek zaglądał do środka, tym bardziej Prosiaczka tam nie było"…

Witaj w nowym, wspaniałym świecie! Świecie AI

O sztucznej inteligencji jest ostatnio bardzo głośno. To nośny temat, nierzadko przedstawiany w tonie sensacyjnym. Czy AI nas zniewoli? Czy wymknie się spod kontroli? A może zabierze nam pracę i zmieni nasze życie w bezproduktywny koszmar? Rzeczywistość wygląda zgoła inaczej, niż zdają się sugerować clickbaitowe nagłówki w prasie i mediach społecznościowych. Sztuczna inteligencja jest obecna w naszym życiu od wielu lat, choć często nie zdajemy sobie z tego sprawy. Służy nam pomocą, gdy szukamy czegoś w sieci, kiedy chcemy coś przetłumaczyć, kupić, porównać albo dotrzeć z punktu A do punktu B. Odsiewa dla nas spam w poczcie internetowej i chroni nasze urządzenia elektroniczne przed cyberatakami. Oczywiście, ma swoje mroczne strony i tych także powinniśmy być świadomi.

Ta książka w przystępny i rzetelny sposób wprowadzi Cię w fascynujący świat sztucznej inteligencji. Bez skomplikowanej matematyki, odwołując się jedynie do niezbędnego minimum na poziomie szkoły średniej, autor przybliża zasady działania i uczenia się sztucznych sieci neuronowych, funkcjonowanie dużych modeli językowych i generatywnej AI, jak słynny ChatGPT, a także omawia ich dzisiejsze możliwości i nadchodzące wyzwania. Odwołuje się przy tym do własnej praktyki zawodowej — od wielu lat współtworzy rozwiązania służące proaktywnej ochronie przed cyberprzemocą i wsparciu osób w kryzysie suicydalnym, dokładając tym samym swoją cegiełkę do prac nad rozwojem sztucznej inteligencji.

Gniewosz **Leliwa**

Innowator i prekursor sztucznej inteligencji w nurcie łączącym uczenie maszynowe, duże modele językowe i wnioskowanie symboliczne. Dyrektor do spraw technologii i współzałożyciel Samurai Labs. Współtwórca rozwiązań AI chroniących przed cyberprzemocą miliony użytkowników internetu na całym świecie. Z wykształcenia fizyk teoretyk, zajmujący się kwantową teorią pola, który porzucił doktorat na rzecz pracy nad sztuczną inteligencją. Autor wielu patentów i publikacji naukowych z obszaru neurosymbolicznej AI, a także jej zastosowania w wykrywaniu i przeciwdziałaniu takim zjawiskom jak cybernękanie, ideacje samobójcze czy *child grooming*.

Helion

helion.pl

HELION S.A.
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-289-2310-2

9 788328 923102

Cena: 59,00 zł