Metodologia tworzenia modeli klasyfikacji

Model klasyfikacji to model, którego wynikiem jest wartość dyskretna (*discrete variable*), czyli taka, która nie ma postaci ciągłej (*continuous variable*). Z kolei wartości ciągłe, to wartości, które można wyrazić w postaci liczb zmiennoprzecinkowych. Wartości dyskretne to wartości zero jedynkowe lub przybierające większą, lecz ograniczoną liczbę postaci wyrażanych liczbami całkowitymi, nazwami, symbolami lub literami.

Większość modeli klasyfikacji nie toleruje zmiennych w postaci szeregu symboli lub liter, dlatego w trakcie przygotowania danych kompiluje się je do postaci liczb całkowitych.

W nowej dziedzinie jaką jest data science większość modeli to modele klasyfikacji kategorycznej. Są to modele odpowiadające na pytania, np. czy będzie przekroczony stan krytyczny czy nie, czy spadnie deszcz czy nie, czy pacjent umrze czy przeżyje.

Nauka z nadzorem

Przeważająca ilość modeli data science wykonywana jest w reżimie nauki z nadzorem (*supervised learning*). Aby stworzyć model ekonometryczny musi istnieć szereg zmiennych niezależnych oraz przynajmniej jedna zmienna zależna, zwana również zmienną wynikową. Zmienne niezależne, zwane też zmiennymi opisującymi, mogą mieć postać ciągłą, jak np.: temperatura, wilgotność, ciężar lub dyskretną, jak dzień tygodnia, płeć, pełna godzina lub pora dnia. Na czym polega nauka z nadzorem?

Wyobraźmy sobie, że mamy już dane, kilkanaście zmiennych opisujących oraz wybraną zmienną wynikową. Nauka z nadzorem polega na tym, że pierwotny zbiór danych dzielimy na treningowy i testowy. Przyjęło się, że zbiór treningowy stanowi ok. 70-80% liczebności zbioru pierwotnego, zaś resztę, czyli 20-30% tego zbioru stanowi zbiór testowy.

Idealnie, gdy pierwotny zbiór danych zostanie podzielony na zbiór testowy i treningowy przy zachowaniu podobnej charakterystyki statystycznej poszczególnych zmiennych opisujących (np. strukturą rozkładu prawdopodobieństwa). Oba zbiory powinny więc być do siebie statystycznie podobne.

Nauka z nadzorem polega na tym, że trenuje się model na zbiorze treningowym, a potem sprawdza się efektywność nauki podstawiając do wytrenowanego modelu zmienne niezależne zestawu testowego. Następnie porównuje się uzyskane z modelu wartości teoretyczne do wartości wynikowych zbioru testowego.

Zależność przyczynowo-skutkowa zmiennych opisujących i wynikowych

Wynik modelu klasyfikacji ma postać dyskretną. Aby model był efektywny musi istnieć istotny związek informacyjny pomiędzy zmiennymi opisującymi a zmienną wynikową. Niestety w przypadku zmiennych dyskretnych nie ma mowy o pomiarze związku przyczynowo-skutkowego za pomocą zwykłej korelacji Pearsona (*Pearson's correlation coefficient*). Gdy zmienne niezależne i zależne mają postać dyskretną (zwaną również postacią kategoryczną) związek przyczynowo-skutkowy pomiędzy nimi mierzony jest za pomocą metody Mutual Information lub Chi-Squared. Gdy zmienne opisujące mają postać ciągłą, ich wpływ na dyskretną zmienną wynikową mierzony jest algorytmem ANOVA lub metodą współzależności Kandella (*Kendall's rank coefficient*).

Ocena jakości klasyfikacji

Załóżmy, że udało nam się zbudować model, który potrafi wskazywać czy ziarno w skupie jest zainfekowane pleśnią czy nie. Na tym etapie skupu pleśń na ziarnach nie jest widoczna. Nasz model podaje nam dwa wskazania:

1: ziarno zainfekowane

0: ziarno niezainfekowane.

Model został wytrenowany na zbiorze treningowym na podstawie kilkunastu dostępnych zmiennych takich jak kolor, masa jednostkowa czy wilgotność. Dzięki nim model potrafi wskazywać, które ziarno wkrótce ulegnie widocznemu zapleśnieniu.

Mając zbiór testowy możemy ocenić jaka jest dokładność klasyfikacji modelu. Wystarczy podstawić do wytrenowanego modelu zmienne niezależne ze zbioru testowego i w ten sposób otrzymać wyniki teoretyczne. Porównując wyniki teoretyczne do laboratoryjnych zawartych w zbiorze testowym możemy szybko wyłapać wszystkie błędy klasyfikacji modelu.

Przy modelach klasyfikacji dychotomicznej istnieją cztery możliwe oceny:

1. *True positive*: model dobrze wskazał ziarno pleśniejące jako ziarno pleśniejące.
2. *True negative*: model dobrze wskazał ziarno zdrowe jako ziarno zdrowe.
3. *False positive*: model wskazał ziarno zdrowe jako ziarno pleśniejące.
4. *False negative*: model wskazał ziarno pleśniejące jako ziarno zdrowe.

Zestawienie czterech stanów klasyfikacji nazywane jest macierzą pomyłek (*confusion matrix*), zwaną również macierzą błędów (*error matrix*). W macierzy tej zlicza się ilość prawidłowych i nieprawidłowych klasyfikacji testowanego modelu.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Źródło: „Measuring Performance: The Confusion Matrix"; GLASS BOX Machine Learning and Medicine, by Rachel Lea Ballantyne Draelos; https://glassboxmedicine.com/2019/02/17/measuring-performance-the-confusion-matrix/

Trudno ocenić jakości klasyfikacji oglądając samą macierz pomyłek. Dlatego wymyślono szereg wskaźników oceny, spośród których najważniejsze są dwa: *recall* oraz *precision*.

Recall to wskaźnik oparty na ilorazie ilości prawidłowo wykrytych spleśnień (klasyfikacje typu: *true positive*) do całego wolumenu spleśniałego ziarna, które przeszło przez model (sumy klasyfikacji typu: *true positive* oraz *false negative*). To po prostu procentowy udział wykrytego spleśniałego ziarna w całkowitej ilości spleśniałego ziarna, które było poddane badaniu przez model. Gdy *recall* wynosi 73% oznacza to, że 27% spleśniałego ziarna została sklasyfikowane jako ziarno zdrowe.

Inaczej działa wskaźnik *precision*. Jest to iloraz dobrze sklasyfikowanego zapleśniałego ziarna (klasyfikacje typu: *true positive*) do całej sumy ziarna, którą model wskazał jako ziarno spleśniałe. *Precision* wskazuje więc jaki jest udział naprawdę zapleśniałego ziarna w ziarnie, które zostało sklasyfikowane przez model jako spleśniałe. Tak więc na przykład wartość *precision* wysokości 82% oznacza, że 18% ziarna zdrowego zostało uznane przez model za ziarno zapleśniałe.

Problem niezbilansowanego zbioru wyników

Model klasyfikacji będzie działał najlepiej, gdy wolumen zapleśniałego ziarna będzie porównywalny do wolumenu ziarna zdrowego. Na szczęście ziarno zapleśniałe jest rzadkością w skupie. Przyjmijmy, że ziarno z niewidocznym zapleśnieniem stanowi zaledwie 1% całego przyjmowanego w skupie ziarna. Istnieje więc ogromna dysproporcja ziarna zdrowego w stosunku do ziarna zarażonego. W takiej sytuacji model nie będzie w stanie dokonać poprawnej klasyfikacji. Kiedy uruchomimy model z łatwością wskaże on, że wszystkie przyjmowane ziarna są zdrowe. Niestety taka jest logika algorytmów matematycznych. Upraszczają one rzeczywistość podejmując decyzje w oparciu o rachunek prawdopodobieństwa. Model, który uzna, że całe ziarno jest zdrowe będzie miał bardzo dobrą względną ocenę jakości klasyfikacji, ponieważ szacunkowo będzie mylił się raz na sto dostaw (przyjęliśmy, że ziarno z zapleśnieniem pojawia się raz na sto dostaw). Jednak wskaźnik *recall* i *precision* pokażą bardzo niską efektywność równą zeru, algorytm ten będzie dla nas bezużyteczny.

Bilansowanie zbioru wyników

Modele klasyfikacji zazwyczaj służą do tropienia zjawisk rzadkich, takich jak malwersacje finansowe, skazy genetyczne, rzadkie choroby lub błędy systemu.

Celem klasyfikacji jest często wyszukiwanie anomalii, których występowanie jest niezwykle unikatowe. Aby model klasyfikacji mógł działać poprawnie musi mieć zbliżoną ilość zjawisk typowych i unikatowych. Jeżeli zjawisk typowych będzie tyle samo co unikatowych, te ostatnie nie będą już unikatowe. Mamy więc uwarunkowanie, które wydaje się być niemożliwe do realizacji. Jednak ten kuriozalny warunek można zrealizować stosując jedną z dwóch technik: nadpróbkowanie (*oversampling*) i podpróbkowanie (*undersampling*).

Technika *oversampling* polega na stworzeniu w zestawie danych treningowych wielkiej ilości sztucznych zdarzeń unikatowych poprzez skopiowanie ich z prawdziwych zdarzeń zarejestrowanych w zestawie treningowym. Ilość zmiennych typowych powinna być podobna. Dzięki temu model nie zatraci ważności żadnej z kategorii.

Technika *undersampling* polega na eliminacji w zestawie danych treningowych populacji zdarzeń typowych tak, aby ich liczba była zbliżona do ilości zdarzeń unikalnych. W ekonometrii króluje centralne twierdzenie graniczne, które wskazuje na powszechne występowanie w przyrodzie rozkładów zbliżonych do rozkładu normalnego.

Gdy liczba obserwacji *n* rośnie do nieskończoności to rozkład zmiennych losowych dąży do rozkładu normalnego. Zachowanie rozkładu normalnego jest warunkiem efektywności większości modeli. Tak więc redukcja liczebności zmiennych w zbiorze treningowym nie wydaje się najlepszym pomysłem. Dlatego też w praktyce badawczej najczęściej wykorzystuje się metodę *oversampling*.

Wykorzystanie modeli klasyfikacji

Ze względu na radykalny charakter wyników modele klasyfikacji są w jakimś stopniu efektywniejsze od modeli regresyjnych, których wyniki mają postać ciągłą. Typowym zastosowaniem modeli regresyjnych są prognozy cen, temperatury, ciśnienia i innych wielkości ciągłych.

Modele regresji mogą na przykład wskazywać prognozę cen na giełdzie, tymczasem modele klasyfikacji kategorycznej mogą tylko mówić czy cena wzrośnie czy spadnie. Modele klasyfikacji mogą również wyznaczyć kilka klas spadków i wzrostów, przez co informacja będzie bardziej precyzyjna. Wydaje się, że rozwój technik predykcyjnych będzie opierał się częściej na przedziałach prognostycznych oferowanych przez modele klasyfikacji niż na wartościach ciągłych oferowanych przez modele regresyjne.

W przeszłości ogromna część analiz statystycznie pokazywała wyniki w postaci przedziałów prawdopodobieństwa. Od tej metody odstąpiono, ponieważ wyniki takie były trudne do interpretacji dla zwykłych ludzi.

Wojciech Moszczyński
Współpraca Ewa Moszczyńska

Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu, specjalista z zakresu ekonometrii, *data science* oraz rachunkowości zarządczej. Specjalizuje się w optymalizacji procesów produkcyjnych i logistycznych. Prowadzi badania w obszarze rozwoju i zastosowania sztucznej inteligencji. Od lat zaangażowany w popularyzację ekonometrii oraz *data science* w środowiskach biznesowych