Data science w piekarni

Czy standaryzacja jest potrzebna przy modelowaniu procesów?

Standaryzacja to jedna z metod przetwarzania surowych danych przygotowywanych do wprowadzenia ich do modeli ekonometrycznych. Celem tego przetwarzania jest dopasowywanie danych do jednej wspólnej skali.

Zgodnie z definicją standaryzacja polega na przetwarzaniu całej populacji w celu uzyskania ujednoliconych parametrów statystycznych, czyli średniej dla populacji równej 0 oraz odchylenia standardowego równego 1.

Dwa typowe błędy w rozumieniu standaryzacji

Standaryzujemy dane wejściowe do modelu tak, aby wszystkie zmienne miały podobną charakterystykę statystyczną. I tu pojawiają się dwa błędy popełniane zwykle przez początkujących badaczy.

Wartości zmiennych mogą przekraczać próg 1 i -1

Po pierwsze, jeżeli odchylenie standardowe jest równe jeden to nie znaczy, że skala zmiennych zatrzymuje się na wartości 1 i -1. To oznacza, że zmienne mogą mieć rozmaite wartości od 0,1 nawet po -2, mimo to suma odchyleń będzie wynosi 1. Pomyłka ta wynika z podobieństwa procesu standaryzacji do popularnego procesu normalizacji. Normalizacja to proces przetwarzania zmiennych wejściowych tak, by wartości zmiennych znajdowały się w przedziale od 0 do 1 lub w innym wypadku od -1 do 1. Przy normalizacji niedozwolone jest, aby zmienna przekraczała próg wartości bezwzględnej 1. W przypadku standaryzacji zmienne mogą swobodnie przekraczać prób wartości, jednak suma odchyleń nie może istotnie odbiegać od wartości 1.

Rozkład prawdopodobieństwa standaryzowanych danych nie musi przypominać rozkładu normalnego

Każdy początkujący analityk danych wie, że typową cechą rozkładu normalnego jest charakterystyka statystyczna opisana jako średnia równa zero i odchylenie standardowe równe 1. Tak się składa, że identyczne parametry ma populacja po przeprowadzeniu jej standaryzacji. Stąd prawdopodobny błąd myślowy, że po standaryzacji otrzymamy piękny rozkład normalny gęstości prawdopodobieństwa.

Standaryzacja opisana jest przez funkcję liniową. Jeżeli jakąś liczbę podstawimy do funkcji liniowej otrzymamy wielokrotność tej wartości. Jeżeli wszystkie wartości z populacji pomnożymy przez 2 lub dodamy 7 to będziemy mieli identyczna strukturę wartości. Na przykład gdy mamy zbiór danych składający się z ciężaru właściwego jednego rodzaju chleba. Robimy wstępną analizę, która mówi, że 20% pieczywa jest za ciężka w stosunku do ustalonej normy, a 45% pieczywa ma niedowagę. Teraz jeżeli te dane podzielimy przez liczbę 15 dalej będziemy mieli zachowaną tą strukturę ciężaru chleba według przyjętych cech: 20/35/45.

Funkcja standaryzacji ma charakter liniowy i zapisywana jest w formie równania: standardized value = $X - \mu / \sigma$, gdzie: μ – średnia populacji przed standaryzacją, σ – odchylenie standardowe populacji przed standaryzacją.

Po co stosuje się standaryzację?

Przyjmijmy, że realizujemy projekt mający na celu poprawę zadowolenia klientów końcowych piekarni. Musimy tak udoskonalić proces wypieku pieczywa, żeby nasi klienci byli bardziej zadowoleni.

Proces wypieku charakteryzuje się wieloma rodzajami zmiennych, które mogą być przydatne do modelowania produkcji. Gromadzimy więc szeregi czasowe opisujące temperaturę wewnątrz pieca, gdzie zakres temperatur wynosi od 90 do 300°C, wilgotność w zakresie 20-60% (zapisywana w danych jako 0,2 do 0,6) oraz ciśnienie wewnątrz rur doprowadzających gorący olej do pieca w zakresie 0,6-3 bary. Dla naszej analizy istotny jest również szereg czasowy temperatury na zewnątrz pieca w zakresie 15-25°C. Mamy też szereg czasowy wartości zużycia gazu w zakresie 0,005-0,01 m^3/min.

Taka różnorodność wartości nominalnych zmiennych nie jest dobra dla procesu modelowania, ponieważ część algorytmów modelowych faworyzuje zmienne o największych wartościach i zaniedbuje wartości stosunkowo niskie. Jest to typowa cecha np. regresji Ridge'a i wielu innych algorytmów. W naszym przypadku, na podstawie własnych domysłów, mamy podejrzenie, że największy wpływ na jakość pieczywa ma temperatura na zewnątrz komory piecowej. Podczas wyciągania chleba z pieca różnica temperatur może spowodować powstanie charakterystycznej skórki, która może okazać się kluczowa dla końcowych klientów. Skórkę chyba każdy lubi, tylko ta na chlebie musi być odpowiednia, nie popękana, nie za cienka. Temperatura na zewnątrz pieca ma dla algorytmu wspomnianej regresji Ridge'a stosunkowo małą wartość w porównaniu z temperaturą wewnątrz pieca. Model może więc potraktować tę kluczową dla skórki temperaturę jako czynnik nieistotny.

Aby uniknąć takiej sytuacji, wszystkie zebrane dane powinny być standaryzowane do tego samego zakresu wartości nominalnej. Do tego właśnie służy standaryzacja.

Standaryzuje się oddzielnie zbiór testowy i zbiór treningowy

Wyjaśnienie powodów takiego podejścia nie jest już takie oczywiste jak w przypadku

wcześniejszych potencjalnych błędów procesu standaryzacji.

Błąd polega na tym, że niektórzy badacze za jednym zamachem standaryzują wszystko, a następnie dzielą populację na zbiór testowy i zbiór treningowy. To oczywisty błąd. Należy najpierw podzielić populację, a potem oddzielnie standaryzować zbiór testowy i zbiór treningowy.

Dlaczego? Ponieważ wspólna standaryzacja tych zbiorów powoduje, że są one podobne, a przecież takiej sytuacji nie ma gdy model pobiera „prawdziwe dane" kiedy już normalnie pracuje po wdrożeniu. Wspólna standaryzacja danych testowych i treningowych prowadzi do swoistego „wycieku danych". Zestandaryzowany zbiór testowy jest wtedy w jakimś stopniu spokrewniony z zestandaryzowanym zbiorem treningowym. To spokrewnienie jest wprawdzie nieistotne statystycznie lecz suma wielu nieistotnych błędów prowadzi do poważnego błędu. Powinniśmy więc robić wszystko, aby unikać drobnych błędów.

Pięć powodów, dla których stosujemy standaryzację

Istnieje pięć powodów, dla których używamy standaryzacji w pracy z modelowaniem lub grupowaniem.

1. Porównywalność skali

Jak wyjaśniłem w przykładzie na początku, nie jest dobrze dla modelowania, jeśli poszczególne cechy znacznie się od siebie różnią. Jest to szczególnie szkodliwe dla niektórych algorytmów bazujących na odległościach, np. Support Vector Machine (SVM). Należy doprowadzić do wspólnej skali cech w określonym przedziale wartości. Ograniczenie wszystkich wartości do rygorystycznej charakterystyki statystycznej oznacza, że wszystkie te dane zostaną zmniejszone według proporcji. Tak więc, jeżeli rozkład populacji był asymetryczny to standaryzacja zachowa tę asymetryczność. Standaryzacja przetwarza dane od wartości wstępnej do wartości znormalizowanej z uwzględnieniem położenia tych wartości względem średniej tej populacji cech. Zatem temperatury wewnątrz pieca i temperatura na zewnątrz urządzenia po standaryzacji dla laika mogą być bardzo podobne, pozostają jednak identyczne z punktu widzenia gęstości prawdopodobieństwa zdarzeń. Ta ujednolicona charakterystyka wartości gwarantuje, że modele będą podobnie traktowały wszystkie cechy procesu. Warto wspomnieć, że jeśli umieścimy zestandaryzowane dane w modelu, również zmienne zależne, możemy spodziewać się, że otrzymamy zestandaryzowane wyniki końcowe. Często jest to koszmarem dla początkujących badaczy. Ponieważ standaryzacja jest funkcją liniową, możemy przeprowadzić odwrotny proces standaryzacji. Możemy powstrzymać się też od standaryzacji wejściowych danych wynikowych, tzn. zmiennych zależnych.

2. Przyspieszenie poszukiwania zbieżności

Algorytmy optymalizacyjne zwykle bazują na punktach ekstremalnych powierzchni lub punktach przecięcia wektorów. Standaryzacja przyspiesza działanie takich algorytmów. Powierzchnia poszukiwań po przeprowadzonej standaryzacji jest bardziej kulista, dzięki czemu algorytmom łatwiej jest znaleźć punkt styku. Bez standaryzacji hiperpowierzchnie są od siebie odległe, co stanowi przeszkodę w procesach obliczeniowych.

3. Standaryzacja łagodzi wartości odstające

Stosując standaryzację eliminujemy wartości odstające. Na przykład, jeśli mamy dwie cechy zmiennych niezależnych, o podobnym zakresie wartości wprowadzonych do modelu regresji liniowej, a jedna z cech ma kilka wartości odstających, wartość wyniku może zostać zniekształcona. W tym przypadku badacz stosujący podejście klasyczne będzie musiał ręcznie eliminować wartości odstające. Jeśli zdecydujemy się na standaryzację lub normalizację obu wskaźników, może nie prowadzić eliminacji wartości odstających.

4. Konieczna jest standaryzacja algorytmów bazujących na odległościach

Algorytmy oparte na odległości są to niektóre algorytmy wyszukiwania podobieństw, na przykład algorytm K-średnich do grupowania i model K-Najbliższych Sąsiadów (K-NN), używany do zastosowań klasyfikacyjnych lub regresyjnych. Jeśli mamy bardzo dużą odległość między średnią wartością zmiennych, algorytmy modeli mają trudności z właściwym dostosowaniem odległości, które odzwierciedlają wynikową jakość obliczeń. Problem ten dotyczy również grupowania algorytmem K-means.

5. Standaryzacja pomaga w interpretacji wielkości wpływu poszczególnych cech na wartość wynikową

Jak pisałem na początku, niektóre modele traktują pewne zmienne lepiej jeżeli ich wartość nominalna jest wysoka. Dzięki temu pewne cechy o większej wartości mają większy wpływ na zmodelowany proces niż ma to miejsce w realnym procesie. Jest to często spotykana dysfunkcja modeli. Kontynuacją tej błędnej interpretacji są błędne decyzje biznesowe. Wróćmy do naszego przykładu, w którym mamy wysoką temperaturę wewnątrz pieca i niską temperaturę na zewnątrz pieca.

Algorytm ważności cechy, np. model Shapley'a oceniający znaczenie zmiennych niezależnych w wynikach regresji, zidentyfikuje, że np. temperatura wewnątrz pieca jest najważniejsza dla poprawy zadowolenia klienta. Pamiętamy, że model może tak pomyśleć, ponieważ ta zmienna jest nominalnie najwyższa z wszystkich zmiennych. Wiemy, że to nieprawda, wiemy, że temperatura na zewnątrz pieca, albo różnica pomiędzy temperaturami na zewnątrz i w środku pieca odgrywa kluczową rolę w jakości skórki pieczywa. Zatem bez standaryzacji popełniamy systematyczny błąd. Jeśli dojdziemy do błędnego wniosku, nie poprawimy jakości chleba i nasze wysiłki spełzną na niczym.

Wojciech Moszczyński

Wojciech Moszczyński – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu, specjalista z zakresu ekonometrii, finansów, *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 *machine learning* oraz *data science* w środowiskach biznesowych.