Uczenie maszynowe: przygotowanie danych – kodowanie zmiennych kategorialnych

Jak już wiemy z poprzednich wpisów w uczeniu maszynowym wyróżniamy następujące rodzaje zmiennych: ilościowe, kategorialne oraz mieszane. I o ile ze zmiennymi ilościowymi (numerycznymi) modele radzą sobie dość dobrze, bo przecież w końcu operują na liczbach, to pozostałe rodzaje zmiennych dla opracowywanych modeli trzeba odpowiednio przygotować. Stąd też w obecnym wpisie przygotowanie danych – kodowanie zmiennych kategorialnych temat przekodowania zmiennych kategorialnych będzie tematem wiodącym.

Z tej serii wpisów poświęconych przygotowaniu danych dla modeli uczenia maszynowego powstały już:

- <https://ekordo.pl/uczenie-maszynowe-przygotowanie-danych-wprowadzenie/>

- <https://ekordo.pl/uczenie-maszynowe-przygotowanie-danych/>

Dla przypomnienia zmienna kategorialna to taka, która pozwala na przeprowadzenie klasyfikacji serii danych za pomocą wartości związanych z określoną jakością czy kategorią. Zmienna kategorialna pozwala sklasyfikować, pogrupować w oparciu o informację jakościową, czyli najczęściej w danych spotkamy zmienną kategorialną z elementami stanowiącymi łańcuch tekstowy. I obecnie naszym zadaniem będzie przekodowanie takiej zmiennej do jej numerycznej reprezentacji, tak by można ją uwzględnić w wielu algorytmach uczenia maszynowego.

Techniki kodowania zmiennej kategorialnej

Kodowanie zmiennej kategorialnej możemy przeprowadzić na wiele sposobów. Wraz z rozwojem technik uczenia maszynowego ta dziedzina nauki uległa również znacznemu rozwojowi [1]. Metodyka nie uległa jednak zmianie. Proces kodowania rozpoczynamy od rozpoznania charakteru zmiennej kategorialnej. Jeżeli zmienna ma charakter porządkowy to wówczas będzie można stosować metody z grupy kodowania porządkowego (ang. Ordinal Encoding, Integer Encoding). W przypadku, gdy zmienna nie ma charakteru porządkowego, i ponadto wymuszenie takiego uporządkowania nie byłoby dobrym rozwiązaniem z punktu widzenia opracowywanego modelu, stosujemy metody polegające na przekształceniu kategorii do postaci binarnego wektora (ang. One-Hot Encoding). Poniżej podano prosty przykład przekodowania zmiennej nominalnej o nazwie kolor.

Kolor

Kolor

Czerwony |1|0|0

Zielony |0|1|0

Niebieski |0|0|1

Oczywiście wprawne oko może zauważyć, że takie podejście generuje nam pewną nadmiarowość w kodowaniu i istnieje możliwość skrócenia ciągu binarnego. A i owszem, ale o tym później. Teraz jeszcze wspomną o zmiennej dychotomicznej, czyli takiej która przyjmuje jedynie dwie wartości. Często jako przykład tego rodzaju zmiennej podawana jest płeć. Wówczas zbiór kategorii takiej zmiennej kodujemy najczęściej jako 0 i 1 lub -1 i 1.

Dane do analizy

Podobnie jak poprzednio pracować będziemy na danych pochodzących ze zbioru opisującego nieruchomości House Prices – Advanced Regression Techniques. Link do zbioru danych [2]

Kodowanie z wykorzystaniem metody Ordinal Encoding

Jak wspomniałem kodowanie porządkowe można uznać za naturalny sposób kodowania dla zmiennych porządkowych. W pakiecie *scikit-learn* w module *preprocessing* znajdziemy klasę *OrdinalEncoder*. Wybierzmy jeszcze jakąś zmienną z naszego zbioru. Na pierwszy rzut oka trudno uznać którąś ze zmiennych kategorialnych jako porządkową. Zatem pomysł będzie taki, aby wyznaczyć średnią cenę nieruchomości i powiązać ją z okolicą (. Obliczona średnia cena nieruchomości będzie wskaźnikiem uporządkowania lokalizacji, którą następnie właśnie przekodujemy. Przy okazji zademonstruje kodowanie z żądanym porządkiem kodowania, gdyż OrdinalEncoder standardowo koduje kategorie alfabetycznie. Zatem do dzieła.

Po wykonaniu kodu otrzymujemy zmienną numeryczną identyfikującą lokalizację nieruchomości w taki sposób, ze najmniejsza liczba (zero) została przypisana lokalizacji z najniższą wartością, a największa (24) lokalizacji gdzie cena nieruchomości jest najwyższa.

Kodowanie typu One-Hot Encoding

Dla zmiennych kategorialnych bez uporządkowania kodowanie numeryczne, a więc określające relację między kategoriami może prowadzić do błędnego działania uczonego modelu. Wymuszenie uporządkowania, tak jak zrobiliśmy to poprzednio też nie jest dobrym rozwiązaniem. Dla takich przypadków stosuje się kodowanie One-Hot Encoding, które w języku polskim określa się jako kodowanie „1 z n”. Istotą jest zastąpienie kategorii zmiennej poprzez ciąg binarny o długości n z jedną „1” w ciągu. Modyfikacją tego kodowania jest wykorzystanie ciągu binarnego o długości n-1, z uwagi na to że ciąg binarny o takiej długości zapewnia pełną informację o kategoriach zmiennej. Kodowanie poprzez ciąg „n-1” określony jest jako kodowanie ze ślepą zmienną (ang. Dummy Variable Encoding) i jest powszechnie zalecany przy wdrożeniach modeli uczenia maszynowego. Jednakże są przypadki, gdy n-elementowy ciąg może być zasadniejszy do użycia. Do nich należy choćby sytuacja, gdy interesuje nas wpływ pojedynczej kategorii lub budujemy modele oparte na drzewach decyzyjnych. Kodowanie One-Hot Encoding oprócz swoich zalet ma też i wady, do których należy rozszerzenie przestrzeni zmiennych przy braku dodania dodatkowych informacji. W przypadku wielu zmiennych kategorialnych czy nawet dużej ilości kategorii w zmiennej ma to znaczenie.

Do przeprowadzenia kodowania „1 z n” w naszym zbiorze danych skorzystam z pakietu *scikit-learn*. W module *preprocessing* znajdziemy klasę *OneHotEncoder.* Wcześniej dla uproszczenia realizacji procesu kodowania wybiorę kilka zmiennych kategorialnych z określoną liczbą kategorii.

W rezultacie trzy zmienne zawierające następujące kategorie zostały przekodowane do:

I na koniec powtórzę uwagę, którą mocno zaakcentowałem w poprzednim wpisie. Podobnie jak operacja uzupełniania brakujących danych, tak i kodowanie zmiennych powinno zostać zrealizowane na zbiorze danych uczących, i potem dopiero przetransponowane na zbiór testowy.

1. <https://en.wikipedia.org/wiki/Categorical_variable>

2. <https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data>

Zmienna dychotomiczna to każda cecha badanej osoby, obiektu lub zjawiska, która przyjmuje jedynie dwie wartości. Najprostszym przykładem takiej zmiennej jest płeć.

Obraz zawierający tekst

Opis wygenerowany automatycznie

<https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/>

<https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/>

<https://towardsdatascience.com/encoding-categorical-variables-one-hot-vs-dummy-encoding-6d5b9c46e2db>

<https://en.wikipedia.org/wiki/Categorical_variable>