Sprawozdanie z projektu na temat uczenia maszynowego

Autorzy: Radosław Radziukiewicz, Julia Skoneczna

Opis rozwiązywanego problemu

Naszym celem było stworzenie własnej implementacji lasu losowego. Celem wspomnianego algorytmu miała być predykcja ilości spożywanego alkoholu na podstawie dostarczonego zbioru danych. Rozwiązanie, ze względu na brak podziału zbioru na zbiór uczący i testujący, miało być oceniane na podstawie walidacji krzyżowej.

Podział pracy

Radosław Radziukiewicz: DecisionTreeClassifier.py, Test.py, main.py
Julia Skoneczna: RandomForest.py, CrossValidation.py, main.py

Wykorzystane narzędzia i biblioteki

Program został napisany w języku Python z wykorzystaniem biblioteki pandas.

Opis zbioru danych

Zbiór danych, na którym przeprowadzaliśmy trening i ewaluację modelu, dostępny jest na platformie Kaggle. Nazywa się on “Student Alcohol Consumption” i składa się z 2 podzbiorów o nazwach student-mat i student-por. Oba zbiory składają się z tych samych 33 kolumn zawierających odpowiedzi studentów na zadawane pytania. Zbiór student-mat liczy 395 rekordów. Zbiór student-por składa się z 649 rekordów. Kolumną, której predykcją planujemy się zająć, jest ta o nazwie ‘Dalc’. Oznacza ona dzienne spożycie alkoholu przez studenta. Zawiera ona dane kategoryczne ponumerowane od 1 do 5 (1 oznacza bardzo niskie spożycie, 5 z kolei oznacza duże spożycie). Przed połączeniem zbiorów danych oceniliśmy ich zawartość pod względem kolumny ‘Dalc’, jednak nie zauważyliśmy istotnych różnic. Procentowy udział poszczególnych klasy wynosił: 1: 69%, 2: 19%: 3: 7% 4 oraz 5: 5% (różnice w obu zbiorach danych występują na częściach dziesiątych procent).
Zbiór danych otrzymany jako złączenia obu pomniejszych zbiorów zawiera zatem klasy w wyżej wymienionej objętości i zawiera 1044 rekordy. W zbiorze nie brakuje żadnych wartości (wszystkie rekordy są uzupełnione).

Wstępna analiza danych i zależności

Aby lepiej zrozumieć zbiór danych na którym pracujemy, przeprowadziliśmy wstępną analizę danych oraz zależności, które mogą między nimi występować.

Analiza danych

Większość respondentów pochodzi z obszarów miejskich (759 odpowiedzi). Pozostałe osoby (285 odpowiedzi) pochodzą z obszarów wiejskich. W zbiorze jest mniej odpowiedzi męskich niż żeńskich (453 do 591). Praktycznie tyle samo osób posiada zajęcia pozaszkolne (516) i nie (528). Dominująca większość respondentów planuje podjąć dalszą edukację (aż 955 osób). Większość osób nie jest w związku (673 osoby). Finalnie możemy zaobserwować, iż studenci spożywają w weekendy więcej alkoholu niż na co dzień (395, 235, 200, 238, 73 odpowiednio dla klas 1, 2, 3, 4, 5. Obserwujemy przesunięcie się tendencji z klasy 1 dla dziennego spożycia do klas 1, 2 oraz 3).

Analiza zależności

Ponieważ zbiór danych zawiera bardzo dużo kolumn opisujących poszczególne cechy danego przypadku, niemożliwe było zbadanie zależności pomiędzy każdą parą. Cechami, których zależności postanowiliśmy sprawdzić były: weekendowe spożycie alkoholu (zakładamy, że jak ktoś spożywa dużo alkoholu w weekend, to istnieje większa szansa iż będzie również spożywał go więcej na co dzień), płeć (zakładamy, że mężczyźni mają większe skłonności do codziennego spożywania alkoholu), aktywności pozaszkolne (zakładamy, że osoby które posiadają aktywności pozaszkolne spożywają mniej alkoholu w dzień roboczy) oraz bycie w związku (zakładamy, że single spożywają więcej alkoholu).

Spożycie weekendowe

Praktycznie wszystkie osoby, które w weekend nie piją alkoholu (odp 1) na co dzień też z niego nie korzystają (391 osób). W grupie osób, które w weekend spożywają więcej alkoholu (odp 2 i 3) również przeważają odpowiedzi niskiego korzystania z alkoholu na co dzień (odp 1 udzieliło odpowiednio 178 i 106). Rośnie jednak znaczenie grupy 2 (odpowiednio 52 i 72 osoby). W grupie osób, która spożywa sporo alkoholu (odp 4 i 5) zaczynają dominować już inne grupy (dla grupy 4 dominuje grupa 2 a dla grupy 5 grupa 2). Poniżej prezentujemy dokładną rozpiskę:

Walc	Dalc	
1	1	391
	2	4
	3	1
	4	1
	5	1
2	1	178
	2	52
	3	2
	4	2
	5	1
3	1	106
	2	72
	3	17
	4	5
4	2	56
	1	43
	3	31
	4	8
5	5	24
	3	18
	2	12
	4	10
	1	9

Wniosek: Istnieje korelacja pomiędzy dziennym a weekendowym spożyciem alkoholu. Nie jest natomiast jasne, jak ta korelacja odnosi się do poszczególnych grup. Widzimy raczej pewną tendencję (ten, kto pije więcej alkoholu w weekend, jest bardziej chętny żeby pić go więcej na co dzień) a nie konkretne zależności pomiędzy grupami (wyjątkiem jest klasa ‘Walc’)

Płeć

Widzimy, iż mężczyźni pomimo że stanowią mniej liczną grupę, przeważają w wyższych klasach spożycia alkoholu. Kobiety natomiast dominują klasę 1 tzn. niewielkie spożycie alkoholu. Poniżej widnieje dokładna rozpiska:

sex	Dalc	
F	1	472
	2	91
	3	16
	4	9
	5	3
M	1	255
	2	105
	3	53
	5	23
	4	17

Wniosek: Istnieje korelacja pomiędzy płcią a dziennym spożyciem alkoholu. W przypadku kobiet możemy raczej zakładać spożycie na poziomie 1-2. W przypadku mężczyzna potrzeba jednak dalszych korelacji aby móc skutecznie przyporządkowywać dane grupy.

Aktywności pozaszkolne

W przypadku aktywności pozaszkolnych nie obserwujemy istotnych różnic pomiędzy grupami. Poniżej widnieje dokładna rozpiska:

activities	Dalc	
no	1	359
	2	105
	3	40
	4	15
	5	9
yes	1	368
	2	91
	3	29
	5	17
	4	11

Wniosek: Nie stwierdzamy bezpośredniej korelacji pomiędzy aktywnościami a dziennym spożyciem alkoholu.

Związek

Pomimo, iż na pierwszy rzut oka może wydawać się, że osoby niebędące w związku mają większe tendencja do udzielenia odpowiedzi 1, 2 lub 3 stwierdzamy, że jest to pochopny wniosek. Warto przypomnieć tutaj, iż w grupach tych występuje znacząca różnica liczebności (singli jest sporo więcej) dlatego wartości absolutne nie są dobrą metryką porównawczą. Finalnie nie obserwujemy znaczących wahań w uzyskanych wynikach. Poniżej prezentujemy dokładną rozpiskę:

romantic	Dalc	
no	1	471
	2	126
	3	53
	4	12
	5	11
yes	1	256
	2	70
	3	16
	5	15
	4	14

Wniosek: nie stwierdzamy bezpośredniej korelacji pomiędzy byciem w związku a dziennym spożyciem alkoholu.

Ogólny wniosek

W trakcie powyższej analizy zależności możemy stwierdzić, iż w zbiorze danych występują korelacje pomiędzy dziennym spożyciem alkoholu a innymi cechami. Korelacje te jednak mogą mieć raczej charakter stwierdzeń niż konkretnych przyporządkowań do grup (wyjątkiem będą grupy ‘Walc’ = 1 oraz ‘Sex’ = F). Mamy nadzieję, stworzony przez nas algorytm odkryje zależności i powiązania, które mogą przyczynić się do poprawnej klasyfikacji. Niestety, obawiamy się iż algorytm może działać raczej jak narzędzie minimalizujące błąd predykcji (zamiast przewidywać klasę 4 odpowie 2 a nie 1, tj wyznacza pewien trend) a nie jako klasyfikator (dokładne przyporządkowanie do odpowiedniej grupy).

Drzewo decyzyjne

W celu poprawnej implementacji lasu losowego musieliśmy stworzyć własną implementację drzewa decyzyjnego. Dostarczona przez nas implementacja korzysta z algorytmu ID3 do konstrukcji drzewa. W wyniku wielu testów poprawności jesteśmy pewni poprawności implementacyjnej (testy zawierały m. in. obserwację poprawności konstrukcji podziałów oraz ocenę dokładności klasyfikacji na zbiorze trenującym, która zawsze wynosiła 100%). Kod implementujący drzewo decyzyjne znajduje się w pliku DecisionTreeClassifier.py.

Las losowy

Las losowy zaimplementowaliśmy w klasyczny sposób, tj. na losowych podzbiorach danych z losowo wybranymi atrybutami utworzyliśmy drzewa, z których składa się las, a następnie w celu uzyskania odpowiedzi wybieramy najczęściej występującą odpowiedź w lesie. Kod źródłowy znajduje się w pliku RandomForest.py.

Eksperymenty

W celu zbadania działania naszego lasu przeprowadziliśmy eksperymenty sprawdzające następujące parametry:

1) Wielkość pojedynczego drzewa w lesie
2) Liczba drzew w lesie
3) Liczba zbiorów, na które dzielone są dane przy walidacji krzyżowej (k).

Poprawnością przewidywania nazywamy procent poprawnie przewidzianych danych na zbiorze testowym.

Po wstępnej analizie danych stwierdziliśmy, że badanie innych parametrów, takich jak inny sposób podziału danych (np. 20% testowe, 80% trenujące) nie przyniosłoby interesujących rezultatów.

Ad. 1

Pozostałe parametry: k = 4, rozmiar lasu = 15

Wielkość drzewa	Poprawność przewidywania [%]
15	68.56
30	68.82
45	69.84

Widzimy, że różnice są niewielkie, jednak zwiększenie wielkości drzewa wpływa na lepszą poprawność przewidywania. Jest to raczej spodziewany wynik – większe drzewo ma szansę podjęcia bardziej zniuansowanych decyzji.

Ad. 2

Pozostałe parametry: k = 4, wielkość drzewa = 45

Liczba drzew w lesie	Poprawność przewidywania [%]
35	70.12
45	70.13
60	70.37

Podobnie jak przy wielkości drzewa, obserwujemy niewielką poprawę, jednak wraz z większą liczbą drzew w lesie widzimy lepszą poprawność. Jest to również spodziewane – jeśli będziemy mieć więcej “niezależnych ekspertów”, czyli drzew, mamy większą szansę na to, że jako całość las udzieli poprawnej odpowiedzi. Jednak niewielka poprawa prawdopodobnie wynika z naszego zbioru danych, co rozwiniemy jeszcze we wnioskach końcowych.

Ad. 3

Pozostałe parametry: wielkość drzewa = 45, rozmiar lasu = 60

K	Poprawność przewidywania [%]
2	70.89

3	71.13
5	70.89

W tym przypadku nie obserwujemy zależności jak w poprzednich eksperymentach, tj. poprawy wyników wraz z większą wartością badanego parametru. Dla k = 3 widzimy najlepszą poprawność.

Wnioski

Widzimy, iż najlepszy rezultat dokładności klasyfikacji, jaki udało nam się uzyskać, to wynik 71.1%. Wynik ten jest zdecydowanie niezadowalający. Skuteczność klasyfikacji jest bowiem niewiele większa niż w przypadku, gdyby stworzony model zawsze zwracał odpowiedź 1 (wtedy dokładność klasyfikacji wyniosłaby około 69%).

Pytaniem, które samo się nasuwa, jest to, czy użyta metryka jest właściwa do oceny modelu. Mogliśmy bowiem użyć metryki dokładności, skuteczności (recall) lub metryki F1. Stwierdzamy jednak, iż metryka dokładności jest wystarczająca. Nie obserwujemy bowiem zwodniczo dużej dokładności (co mogłoby sugerować wysoką skuteczność modelu) lecz jego mierne działanie. W tym wypadku stosowanie innych metryk dałoby tak samo niski wynik (a w zasadzie nawet niższy niż ten uzyskany dla metryki dokładności, ponieważ wytrenowane lasy w znacznej większości przypadków przewidywały odpowiedź równą 1).

Niska skuteczność klasyfikacji może wynikać z postawionej wcześniej hipotezy odnośnie zbioru danych. Zakładaliśmy bowiem, iż pomimo występowania pewnych trendów i zależności pomiędzy cechami, dokładna klasyfikacja próbki do konkretnej grupy może być zadaniem trudnym.

Ponadto, w zbiorze danych możemy zaobserwować dominację pewnych grup (odpowiedzi 1 oraz 2 stanowią 88% zbioru danych, gdzie sama klasa 1 stanowi aż 69%), co z kolei zmniejsza zdolności lasu do dokładnej predykcji (statystycznie aż około 70% przypadków w pojedynczym drzewie budującym las stanowią te związane z odpowiedzią 1). Przy założeniu, iż wytrenowane drzewo “nie ma pojęcia co robi” (tzn. zapytania przeprowadzane w drzewie tak naprawdę wykonywane są na cechach mających marginalne znaczenie dla uzyskania odpowiedzi lub drzewo ma wysokie właściwości przetrenowania), to z dużym prawdopodobieństwem można założyć, iż udzieli ono odpowiedzi 1. Takie drzewa mogą stanowić poważny problem w procesie klasyfikacji przykładu przez las (zakładamy bowiem, iż stworzenie drzewa “dobrego” nie jest zadaniem trywialnym, a zatem ich “wiedza ekspercka” będzie zagłuszana przez drzewa udzielające odpowiedzi równej 1).

Kolejnym uzasadnieniem niskiej skuteczności może być fakt, iż las losowy (oraz drzewa decyzyjne) nie są w stanie tworzyć “nowych” cech wynikających z tych już dostarczonych. Być może użycie klasyfikatora w postaci sieci neuronowej pozwoliłoby na odkrycie bardziej złożonych zależności, które z kolei wpłynęłyby na poprawę wyników klasyfikacji.

Podsumowując stwierdzamy, iż niskie wyniki uzyskane przez klasyfikator lasu losowego mogą wynikać ze specyfiki zbioru danych. Podejrzewamy, że w wyniku przeprowadzenia procesów związanych z rozwijaniem zbioru danych (tworzenie nowych cech na podstawie istniejących, rozszerzenie zbioru o nowe przypadki etc.) możliwa byłaby poprawa jakości działania klasyfikatora.

Dzięki temu projektowi nauczyliśmy się, jak w praktyce zaimplementować drzewo decyzyjne oraz las losowy, a także uświadomiliśmy sobie, że jakość działania lasu losowego zależy w dużej mierze od danych, na których chcemy go używać.