WSI Decision Tree Classifier sprawozdanie

Małgorzata Grzanka

05.05.2024

Spis treści

1 Wstep

Celem zadania była implementacja drzewa decyzyjnego tworzonego algorytmem ID3 z ograniczeniem maksymalnej głebokości drzewa. Do analizy działania stworzonego modelu wykorzystam zbiór danych Cardio Vascular Disease Detection z kaggle, gdzie polem do predykcji jest pole *cardio*. Dodatkowym elementem zadania było stworzenie skryptu przygotowujacego dane dla modelu.

2 Skrypt przygotowujace dane dla modelu

2.1 Zbiór danych

Zadany zbiór danych zawierał nastepujace pola:

- wiek (age) : int (days)
- wzrost (height) : int (cm)
- płeć (gender) : 1: first gender, 2: second gender
- ciśnienie skurczowe (ap_hi) : int
- ciśnienie rozkurczowe (ap_lo) : int
- cholesterol (cholesterol) : 1: normal, 2: above normal, 3: well above normal
- glokaza (gluc) : 1: normal, 2: above normal, 3: well above normal
- palenie (smoke) : binary (0/1)
- spożywanie alkoholu (alco) : binary (0/1)
- aktywność fizyczna (active) : binary (0/1)

Przedstawione pola to cechy pacjenta, a zadaniem modelu jest przewidzenie na ich podstawie, czy pacjent ma chorobe serca, czy nie (pole cardio). Mamy zatem do czynienia z klasyfikacja binarna. Dane przechowywane sa w pliku csv.

2.2 Zasada działania skryptu

Skrypt spełnie powyższe cechy:

- Eliminuje niepoprawne dane ze zbioru danych (NaNs albo ujemne, kiedy nie powinno być żadnych ujemnych - np. waga):
  - Usuwa cała kolumne danych, jeśli jej ilość niepoprawnych danych jest wieksza niż jakiś zadany *drop_column_percentage* wszystkich danych.
  - Wypełnia niepoprawne dane średnia z danej kolumny, jeśli ilość jej niepoprawnych danych jest mniejsza niż jakiś zadany *drop_column_percentage*, ale wieksza niż inny, mniejszy zadany *fill_column_percentage*.
  - Usuwa wszystkie próbki z niepoprawnymi danymi w tej kolumnie, jeśli jest ich mniej niż jakiś zadany *fill_column_percentage*.
- Eliminuje outliers'ów (odstajace dane) ze zbioru danych. Jako outliers'y traktowane sa dane które sa wiecej niż 4 odchylenia standardowe od średniej wartości w kolumnie.

3 Sposób działania drzewa decyzyjnego ID3

3.1 Ogólna zasada działania

Drzewo decyzyjne służy do klasyfikacji binarnej, czyli do przewidywania wartości binarnych na podstawie zadanych cech (kolumn/pól) ze zbioru danych. Działa na zasadzie wielokrotnego dzielenia zbioru danych na mniejsze podzbiory na podstawie obecnych w zbiorze cech. Wybierane sa te cechy, które po podzieleniu zapewnia najlepsza pewność w przypisaniu określonej przewidywanej wartości do próbki, która znalazła sie w tym podziorze.

3.2 Dzielenie danych

Dzielenie danych odbywa sie za pomoca zadania pytania Tak / Nie. W przypadku dzielenia po kolumnie zawierajacej dane binarne, jest to bardzo proste. Dane, których wartość w tej kolumnie wynosi 1 zostaja przydzielone do "lewego"podprzedziału, a dane, których wartość w tej kolumnie wynosi 0, zostaja przydzielone do "prawego"podprzedziału. W przypadku danych ciagłych, należy dokonać zamienienia danych na dane binarne. W moim modelu polega to na rozpatrzeniu warunku, czy wartość danej próbki w tej kolumnie jest mniejsza lub równa jakiejś zadanej wartości (wartość ta jest wzieta z określonym krokiem z wszystkich możliwych wartości w tej kolumnie). Nastepnie podział nastepuje podobnie jak przy danych binarnych - jeśli warunek jest spełniony, próbka zostaje przydzielona do "lewego"podprzedziału, a jeśli nie - do "prawego".

3.3 Wybór kolumny, po której sie dzieli

Cecha, po której dzielony jest zbiór danych lub jakiś jego kolejny podzbiór, powinina zapewniać najlepsza pewność w przypisaniu określonej przewidywanej wartości do próbki, która znalazła sie w tym podziorze. Aby znaleźć ta ceche, można zastosować na przykład entropie lub gini impurity. W moim modelu zdecydowałam sie na ta druga opcje. Dla każdego możliwego podziału liczone jest gini impurity i wybierane jest te, które ma najmniejsza wartość (nieczystość podziału jest możliwe najmniejsza).

3.4 Kryterium stopu

Dzielenie zbioru na kolejne podzbioru w moim algorytmie kończy sie, gdy:

- podzbiór osiagnie jakaś zadana głebokość
- liczba próbek w podzbiorze jest mniejsza, niż zadana maksymalna liczba próbek
- wszystkie próbki w podzbiorze sa jednej klasy (tzw. liść czysty)

W dalszej analizie skupie sie jedynie na wpływie maksymalnej głebokość na dokładność precyzji, traktujac pozostałe hiperparametry jako zmienne stałe.

4 Hiperparametry

Mój model przyjmuje nastepujace hiperparametry:

- maksymalna głebokość - maksymalna ilość podzbiorów w jednej gałezi
- minimalna liczba próbek - ile minimalnie próbek moze byc w podzbiorze (ustawiona na stałe na 1)
- limit wartości dyskretnych - ile maksymalnie moze byc różnych wartości dyskretnych w kolumnie numerycznej (stosowana, aby przysieszyć trenowanie modelu)
- krok podziału wartości dyskretnych - jeśli liczba unikalnych wartości dyskretnych jest wieksza niż limit, do warunków podziału brane sa wartości generowane z zadanym krokiem
- kryterium oceny - może być gini albo entropia (entropia niestety jeszcze nie zaimplementowana)

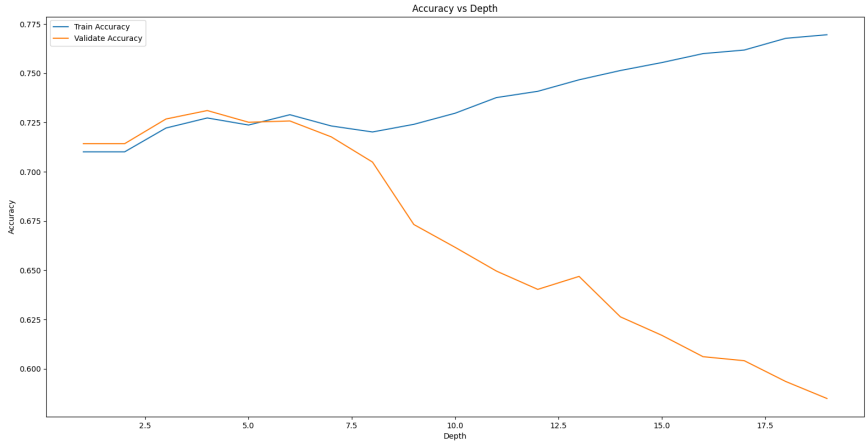
5 Wpływ głebokości na dokładność predykcji

Zbadam wpływ głebokości na dokładność predykcji modelu. W tym celu, minimalna liczba próbek w podzbiorze zostanie ustawiona na 1, aby rozpatrywać jedynie sam wpływ głebokości. Bez zmian pozostaja także hiperparametry: limit wartosci dyskretnych (300), krok podzialu warto sci dyskretnych (100) oraz kryterium oceny (gini).

5.1 Sposób badania wpływu

Dla zbadania wpływu głebokości, podzieliłam zbiór danych na 3 podzbiory - zbiór treningowy, zbiór walidacyjny i zbiór testowy. Model wytrenuje na zbiorze treningowym z różnymi głebokościami w przedziale 1-19, a nastepnie użyje go do predykcji zbiorów treningowego oraz walidacyjnego, aby znaleźć optymalna głebokość. Nastepnie, dokonam ostatecznej predykcji na zbiorze testowym.

5.2 Wyniki



Rysunek 1: Wpływ głębokości na dokładność predykcji zbioru treningowego i walidacyjnego

Jak widać na powyższym wykresie, dokładność predykcji dla zbioru walidacyjnego rośnie do głębokości równej około 4 - wtedy osiaga najwieksza wartość równa około 73%. Nastepnie maleje aż do wartości poniżej 60% skuteczności dla głębokości 19.

W przypadku danych treningowych, dokładność predykcji rośnie w nim wraz ze wzrostem głębokości modelu - od wartości 71% do ponad 76%.

```
->->-> Model rating for depth 1<-<-<-
Confution matrix values (tp, fp, fn, tn): (5094, 1665, 3213, 7036)
Model accuracy: 0.7131937911571026
Precision: 0.7536617842876165
Recall: 0.6132177681473456
F-measure: 0.6762246117084827
FPR: 0.19135731525112057
TNR: 0.8086426847488795
```

Rysunek 2: Wyniki dla danych testowych dla głębokości 1

```
->->-> Model rating for depth 4<-<-<-
Confution matrix values (tp, fp, fn, tn): (5918, 2283, 2389, 6418)
Model accuracy: 0.7253057384760113
Precision: 0.7216193147177173
Recall: 0.712411219453473
F-measure: 0.7169857039011388
FPR: 0.26238363406505
TNR: 0.7376163659349501
```

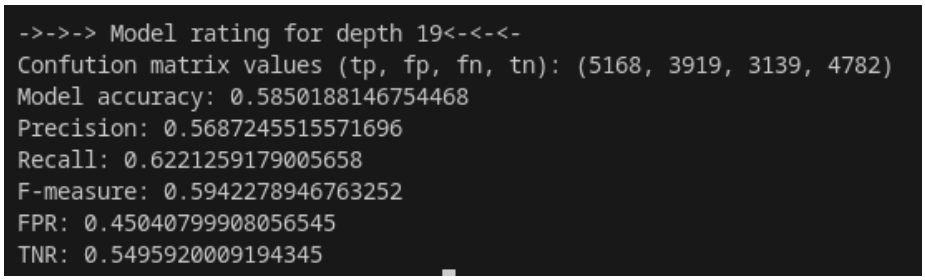
Rysunek 3: Wyniki dla danych testowych dla głębokości 4

```
->->-> Model rating for depth 5<-<-<-
Confution matrix values (tp, fp, fn, tn): (5303, 1700, 3004, 7001)
Model accuracy: 0.7234242709313264
Precision: 0.757246894188205
Recall: 0.6383772721800891
F-measure: 0.692749836708034
FPR: 0.1953798413975405
TNR: 0.8046201586024595
```

Rysunek 4: Wyniki dla danych testowych dla głębokości 5

```
->->-> Model rating for depth 8<-<-<-
Confution matrix values (tp, fp, fn, tn): (6124, 2959, 2183, 5742)
Model accuracy: 0.6976716839134525
Precision: 0.6742265771220962
Recall: 0.7372095822800048
F-measure: 0.7043128234617596
FPR: 0.34007585335018964
TNR: 0.6599241466498104
```

Rysunek 5: Wyniki dla danych testowych dla głębokości 8



Rysunek 6: Wyniki dla danych testowych dla głebokości 19

5.3 Wnioski

5.3.1 Wpływ głebokości na dokładność modelu

Spadek dokładności dla danych walidacyjnych,a wzrost dokładności dla danych treningowych przy wzroście maksymalnej głebokości wiaże sie z zjawiskiem przeuczenia (over-fitting). Przez zbyt dokładne podzielenie drzewa poprzez dane treningowe, model nie oddaje ogólnego trendu w korelacji miedzy danymi, a jest jedynie dobrym odwzorowaniem podziału danych testowych. Z tego powodu, predykcje na danych testowych sa dobre, ponieważ model dobrze sie do nich dopasowuje. Natomaist przez to, że ogólny trend nie jest w nim uwzgledniony, predykcje próbek innych niż treningowe okazuja sie bardzo czesto niepoprawne (dane walidacyjne).

Wzrost dokładności zarówno dla danych walidacyjnych, jak i treningowych przy mniejszych dokładnościach wiaże sie z niedouczeniem (under-fitting). Dane sa zbyt mało podzielone, aby oddawać trend w korelacji miedzy cechami, a przewidywana klasa, wiec zarówno predykcje dla danych walidacyjnych, jak i testowych, sa niepoprawne.

Za optymalna wartość głebokości uznaje wartości z okolicy punktu maksymalnego wykresu danych walidacyjnych. Sa to wartości 3, 4 oraz 5, z czego 4 daje najlepsza możliwa predykcje dla danych testowych - 0.725%.

5.3.2 Wpływ głebokości na pozostałe elementy oceny modelu

Do pozostałych elementów oceny modelu zaliczamy:

- precision - poprawnie przewidziane jedynki / wszystkich przewidzianych jedynek (okładność pozytywnych przewidywań modelu).
- recall - poprawnie przewidziane jedynki / wszystkie jedynki w danych (zdolność do przewidywania jedynek)
- true negative rate (TNR) - poprawnie przewidziane zera / wszystkie zera w danych (zdolność przewidywania zer)

Jak widać w zamieszczonych wyżej wynikach, wzrost maksymalnej głebokości podczas trenowania modelu spowodował spadek TNR oraz Precision oraz wzrost Recall. Maksymalne wartości Precision oraz TNR i minimalne Recall osiagane sa dla wartości z przeciecia sie krzywych dokładności dla danych walidacyjnych i treningowych (głebokości 5 oraz 6). Jednak dla bardzo dużych głebokości (19), przeuczenie jest zbyt duże i wszytskie elementy sa negatywne.

Ze wzgledu na wzrost dopasowania model zaczał przewidywać wiecej jedynek, przez co procent dobrze przewidzianych jedynek wzrósł (Recall). Stało sie to jednak kosztem spadku dokładności w przewidywaniu zer (TNR) i wzrostu liczby niepoprawnie przewidzianych jedynek (Precision). Inaczej mówiac, zbyt duża ilość przewidzianych jedynek sprawiła, że wiele z nich zostało przewidzianych niepoprawnie, a liczba przewidzianych zer spadła, co zmiejszyło dokładność przewidywania zer.