WSI - ID3

Jan Szymczak

Maj 2024

1 Cel ćwiczenia i opis problemu

1.1 Opis problemu i wybrane podejście

Celem ćwiczenia jest implementacja drzew decyzyjnych z użyciem algorytmu ID3 oraz ograniczaniem maksymalnej głębokości drzewa. Następnie, algorytm został wykorzystany do stworzenia modelu na podstawie danych Cardio Vascular Disease Detection ze strony: strona z danymi. Dane zostały podzielone na zbiór uczący, walidacyjny i testujący z pomocą biblioteki *sklearn*.
W przypadku osiągnięcia maksymalnej głębokości, albo w przypadku, gdy nie istnieje podział następnego klasyfikatora, algorytm zwraca klasę o największej częstości. Ponadto, zarówno w metodzie *fit*, która służy do uczenia modelu, jak i w metodzie *predict* służącej do przewidywania wyniku na podstawie modelu, sprawdzana jest poprawność danych (to, czy wszystkie klasy są wartościami dyskretnymi). W przypadku, gdy tak nie jest, zdecydowałem się na wyliczenie średniej arytmetycznej ze wszystkich wartości danej kolumny oraz zaokrąglenia wyniku oczywiście do jedności, aby utrzymać dyskretyzacje. Co prawda, te podejście powoduje, że reszta klas staje się bardziej znacząca, dokładniej $1 + n$ znacząca, gdzie n oznacza ilość próbek, jednak w przypadku działań na danych o dużej wielkości nie ma to większego znaczenia.

1.2 Dyskretyzacja

Dyskretyzację przeprowadziłem następująco:

- Wiek (po przeliczeniu wartości na lata, w danych wiek podany jest w dniach):
  - 1: zakres od 0 do 35 lat
  - 2: zakres od 35 do 45 lat
  - 3: zakres od 45 do 65 lat (w danych nie znajdują się starsi pacjenci)
- Waga i wzrost:
  dyskretyzację tych danych przeprowadziłem pośrednio. Uznałem, że w przypadku choroby serca ani wzrost, ani waga same w sobie nie dają dużo informacji - często z większym wzrostem idzie w parze także i większa waga. Dlatego na podstawie tych danych obliczyłem wskaźnik BMI i zastąpiłem wymienione wcześniej dwie kolumny jedną z BMI, o dyskretnych wartościach według:
  - 1: 0-24.9
  - 2: 24.9 - 29.9
  - 3: 29.9 - 39.9
- Ciśnienie skurczowe:
  - 1: 0 - 120
  - 2: 120 - 140
  - 3: 140 - 190
- Ciśnienie rozkurczowe:
  - 1: 70 - 80
  - 2: 0 - 70 i 80 - 90 (zbyt niskie ciśnienie rozkurczowe także jest oznaką chorób serca)
  - 3: 90 - 120

1.3 Parametry

Jedynym parametrem jest w tym przypadku głębokość drzewa. Zbyt niska prowadzi do słabego wytrenowania drzewa, ale przyspiesza działanie algorytmu. Zbyt wysoka jednak powoduje przeuczenie algorytmu, który za bardzo dopasowuje się do zbioru danych uczących, w których osiąga świetną dokładność, tracąc bardzo na dokładności w zbiorze testowym i wydłużając czas działania programu.

2 Wyniki

Po stworzeniu implementacji algorytmu przystąpiłem do szukania optymalnej wartości hiperparametru *depth* odpowiadającemu głębokości drzewa. Model został nauczony na zbiorze danych uczących, a następnie jego wyniki były sprawdzane na zbiorze danych walidacyjnych. Tak przedstawia się wykres wyników dla zakresu głębokości 1 - 10:

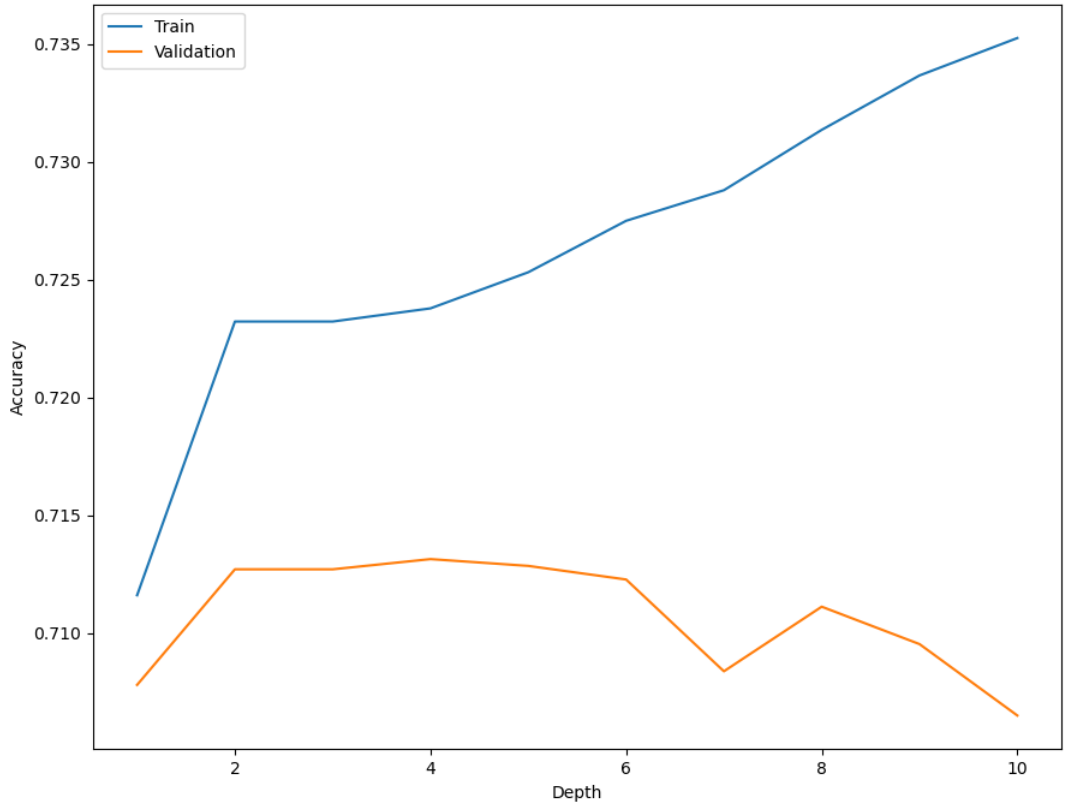

Figure 1: Wykres przedstawiający dokładność dopasowania danych uczących i trenujących w zależności od głębokości drzewa

Dokładność dla danych uczących stale rośnie i ma to jak najbardziej sens, drzewo wraz ze wzrostem głębokości będzie coraz lepiej dopasowywało się do nich. Z kolei w przypadku dokładności dla danych walidacyjnych, najlepsza wartość osiągana jest dla głębokości 4. Dalej (nie licząc przypadkowego wzrostu dla głębokości 8) następuję stały spadek dokładności, mamy powoli do czynienia ze wspomnianym wcześniej przeuczaniem.
Po uzyskaniu optymalnej głębokości dla wybranych zbiorów danych, przeprowadziłem dopasowanie dla zbioru danych testujących. Ostatecznie, dokładności wynosiły:

- Dla zbioru uczącego: 72.4%
- Dla zbioru walidacyjnego: 71.3%
- Dla zbioru testującego: 73.0%

W napisanym przeze mnie ostatecznym teście (metoda *final_test*), znajduje się także wywołanie metody *print_tree*, która wypisuje w terminalu uzyskane drzewo. Klasą o najwyższej zdobyczy informacyjnej na samym początku okazało się ciśnienie skurczowe, *ap_hi*. Następnie, dla wartości *ap_hi* 1 i 2 był to cholesterol, dla wartości 3 był to poziom glukozy. Dalej, klasy już są różnorodne, co można zobaczyć po wywołaniu podanej metody.

3 Komentarz

W przypadku tego algorytmu kluczowa jest wybrana dyskretyzacja danych. To ona w dużej mierze wpływa na to, jak dokładne wyniki otrzymamy. W omawianym przykładzie odpowiednia dyskretyzacja nie jest oczywista, wiąże się ona z pewną wiedzą medyczną, aby móc dobrać odpowiednie wartości dla danych zakresów. Mimo to, procent dokładności i tak jest wysoki, a algorytm pomimo dużej ilości pacjentów w zbiorze danych, zwraca wyniki w zadowalającym czasie.