SI - SPRAWOZDANIE LAB NR 8

Maciej Budzowski, dzienne, grupa L5 31/05/2020

Link do pliku .tex na platformie OverLeaf: OVERLEAF

Zad. 1: Co to jest drzewo decyzyjne? (klasyfikator)

Odpowiedź:

Jest to wielostopniowy system decyzyjny, w którym klasy są kolejno odrzucane, dopóki nie osiągniemy ostatecznie przyjętej klasy.

Podział elementów:

- Drzewo decyzyjne jest formą opisu wiedzy klasyfikującej
- Węzłom drzewa odpowiadają atrybuty eksplorowanej relacji
- Krawędzie opisują wartości atrybutów
- Liśćmi drzewa są wartości atrybutu klasyfikacyjnego

Przestrzeń cech jest podzielona na unikalne regiony, odpowiadające klasom, w sposób sekwencyjny.

Sekwencja decyzji jest stosowana do poszczególnych elementów, a pytania, na które należy odpowiedzieć, mają postać „czy cecha xi <= a?” gdzie a jest wartością progową.

Takie drzewa są znane jako zwykłe binarne drzewa klasyfikacji (OBCT)

Zad. 2: Na jakiej zasadzie działa klasyfikator Random Forest?

Odpowiedź:

Algorytm Losowych Lasów Drzew (ang. Random Forest) - jego idea polega na stworzeniu grupy dużej liczby pojedynczych drzew decyzyjnych, które działają jako zespół. Każde pojedyncze drzewo w losowym lesie wyrzuca prognozę klasy, a klasa z największą liczbą głosów staje się prognozą naszego modelu.

Podstawowa koncepcja algorytmu *Random Forest* jest prosta, ale potężna - mądrość tłumów. Podczas gdy niektóre drzewa mogą się mylić, wiele innych drzew będzie miało rację, więc jako grupa drzewa mogą poruszać się we właściwym kierunku.

W odróżnieniu od klasycznych drzew decyzji, losowe drzewa budowane są na tej zasadzie, iż podzbiór analizowanych cech w węźle dobierany jest losowo.

Zad. 3: Analiza statystyczna wybranego zbioru danych

Użyty zbiór danych

Zbiór danych wybrałem z listy z linku LINK Z POLECENIA.

Do wykonania zadania użyłem zbioru ***Statlog (Australian Credit Approval)***, linki do źródeł zamieszczam poniżej:
DANE (PLIK) (link prowadzi do oryginalnego pliku)
DANE (STRONA) (link prowadzi do strony datasetu)

Opis/temat zbioru: Wedle strony datasetu: “Ten plik dotyczy aplikacji kart kredytowych. Wszystkie nazwy i wartości atrybutów zostały zmienione na bezsensowne symbole w celu ochrony poufności danych.”

Autor:
poufny, przesłane przez quinlan ’@’ cs.su.oz.au

Parametry użytego zbioru:
- liczba obserwacji - 690
- liczba cech, atrybutów - 14 (+1 - klasa)
- liczba klas - 2 (0, 1 - odpowiednio "-", "+")
- liczba elementów na klasę - "-": 383, "+": 307

Rodzaje cech w zbiorze:
Wedle strony: “Istnieje 6 atrybutów liczbowych i 8 atrybutów kategorycznych. Etykiety zostały zmienione dla wygody algorytmów statystycznych. Na przykład atrybut 4 pierwotnie miał 3 etykiety p, g, gg i zostały one zmienione na etykiety 1,2,3.”

Lista cech w zbiorze (wedle strony):

- A1: 0,1 kategoryczne (poprzednio: a,b)

- A2: liczbowe
- A3: liczbowe
- A4: 1,2,3 kategoryczne (poprzednio: p,g,gg)
- A5: 1,2,3,4,5,6,7,8,9,10,11,12,13,14 kategoryczne (poprzednio: ff,d,i,k,j,aa,m,c,w,e,q,r,cc,x)
- A6: 1,2,3,4,5,6,7,8,9 kategoryczne (poprzednio: ff,dd,j,bb,v,n,o,h,z)
- A7: liczbowe
- A8: 1, 0 kategoryczne (poprzednio: t,f)
- A9: 1, 0 kategoryczne (poprzednio: t,f)
- A10: liczbowe
- A11: 1, 0 kategoryczne (poprzednio t,f)
- A12: 1, 2, 3 kategoryczne (poprzednio: s,g,p)
- A13: liczbowe
- A14: liczbowe
- A15: 1,2 atrybut klasy (poprzednio: +,-)

Przygotowanie danych

W ramach przygotowania danych do pracy nie musiałem dokonywać zmian bezpośrednio w oryginalnym pliku. Natomiast zgodnie z metodą z odbytych ćwiczeń "wyciąłem" z wczytanego pliku kolumny zawierające dane binarne. Dzieje się to w trakcie wykonywania skryptu.

Kod

Link do projektu ze skryptem: GITHUB

Model klasyfikacyjny stworzyłem przy użyciu skryptu języka R. Opiera się on na kodzie z odbytych ćwiczeń.

Proces przebiega następująco (w kolejności):

- Wczytujemy plik z danymi.
- Pozbywamy się kolumn z danymi binarnymi.
- Dobieramy zbiory treningowe.
- Uruchamiamy algorytm
- Obliczamy macierz pomyłek
- Obliczamy recognition rate na podstawie macierzy pomyłek

Obliczenia wykonywały się szybko, więc w tym przypadku nie było potrzeby automatyzacji procesu.

Proces wykonałem (tak jak na ćwiczeniach) w kilku próbach, przy użyciu różnych wielkości zbioru treningowego. Wielkości starałem się dobierać proporcjonalnie do tego, jak to się odbywało na ćwiczeniach.

Wyniki

Wielkość zbioru treningowego: 5

	0	1	err
0	**0.9815**	0.0185	0.0185
1	0.7418	**0.2582**	0.7418

Tablica 1: *macierz pomyłek przy zbiorze treningowym równym 5 próbek*

Recognition Rate: 0.6584

Wielkość zbioru treningowego: 10

	0	1	err
0	**0.7037**	0.2963	0.2963
1	0.3245	**0.6755**	0.3245

Tablica 2: *macierz pomyłek przy zbiorze treningowym równym 10 próbek*

Recognition Rate: 0.6912

Wielkość zbioru treningowego: 15

	0	1	err
0	**0.8880**	0.1120	0.1120
1	0.4633	**0.5367**	0.4633

Tablica 3: *macierz pomyłek przy zbiorze treningowym równym 15 próbek*

Recognition Rate: 0.7319

Wielkość zbioru treningowego: 30

	0	1	err
0	**0.8934**	0.1066	0.1066
1	0.4014	**0.5986**	0.4014

Tablica 4: *macierz pomyłek przy zbiorze treningowym równym 30 próbek*

Recognition Rate: 0.7621

Wielkość zbioru treningowego: 60

	0	1	err
0	**0.8470**	0.1530	0.1530
1	0.3141	**0.6859**	0.3141

Tablica 5: *macierz pomyłek przy zbiorze treningowym równym 60 próbek*

Recognition Rate: 0.7762

Wielkość zbioru treningowego: 100

	0	1	err
0	**0.8494**	0.1506	0.1506
1	0.2481	**0.7519**	0.2481

Tablica 6: *macierz pomyłek przy zbiorze treningowym równym 100 próbek*

Recognition Rate: 0.8068

Wnioski

Wyniki sklasyfikowania już dla 5 próbek treningowych osiągnęły wynik 66%.

Procent prawidłowo sklasyfikowanych obiektów wzrastał wraz z zwiększaniem wielkości zbioru treningowego.

Na podstawie osiągniętych wyników wnioskuję, że wielkość zbioru treningowego ma znaczący wpływ na osiągane wyniki w klasyfikacji metodą ***Random Forest***.