Wprowadzenie do sztucznej inteligencji - ćwiczenie 6

Igor Kraszewski
310164

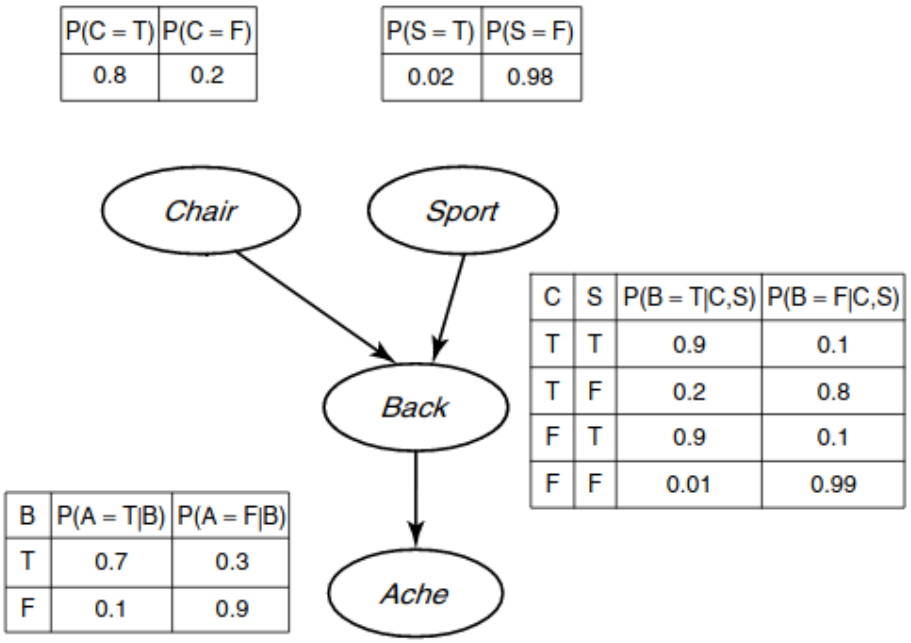
Warszawa, Styczeń 2023

Spis treści

1. Zadanie 2
2. Wyniki 3

1. Zadanie

Proszę zaimplementować losowy generator danych, który działa zgodnie z rozkładem reprezentowanym przez daną sieci bayesowską.



P(C = T)	P(C = F)
0.8	0.2

P(S = T)	P(S = F)
0.02	0.98

C	S	P(B = T|C,S)	P(B = F|C,S)
T	T	0.9	0.1
T	F	0.2	0.8
F	T	0.9	0.1
F	F	0.01	0.99

B	P(A = T|B)	P(A = F|B)
T	0.7	0.3
F	0.1	0.9

Rys. 1.1.

Sieć ta opisuje zależności między (zero-jedynkowymi) zmiennymi losowymi i dana jest w postaci opisu grafu połączeń oraz tabel prawdopodobieństw warunkowych. Wejście algorytmu: ile przykładów wygenerować, opis struktury prostej sieci (według własnego formatu) oraz tabele prawdopodobieństw należy wczytać z pliku tekstowego. Wyjście: plik tekstowy z przykładami. Strukturę sieci i tabele prawdopodobieństw widać na rysunku. Klasa to „Ache” (czy bolą plecy), pozostałe węzły to atrybuty („Back” to uszkodzenie kręgosłupa (drobne, czasem nie skutkujące bólem)). Wytworzony zbiór podzielić i użyć do treningu i testowania klasyfikatora utworzonego na wcześniejszych ćwiczeniach. Jakie uzyskujemy wyniki? Wnioski?

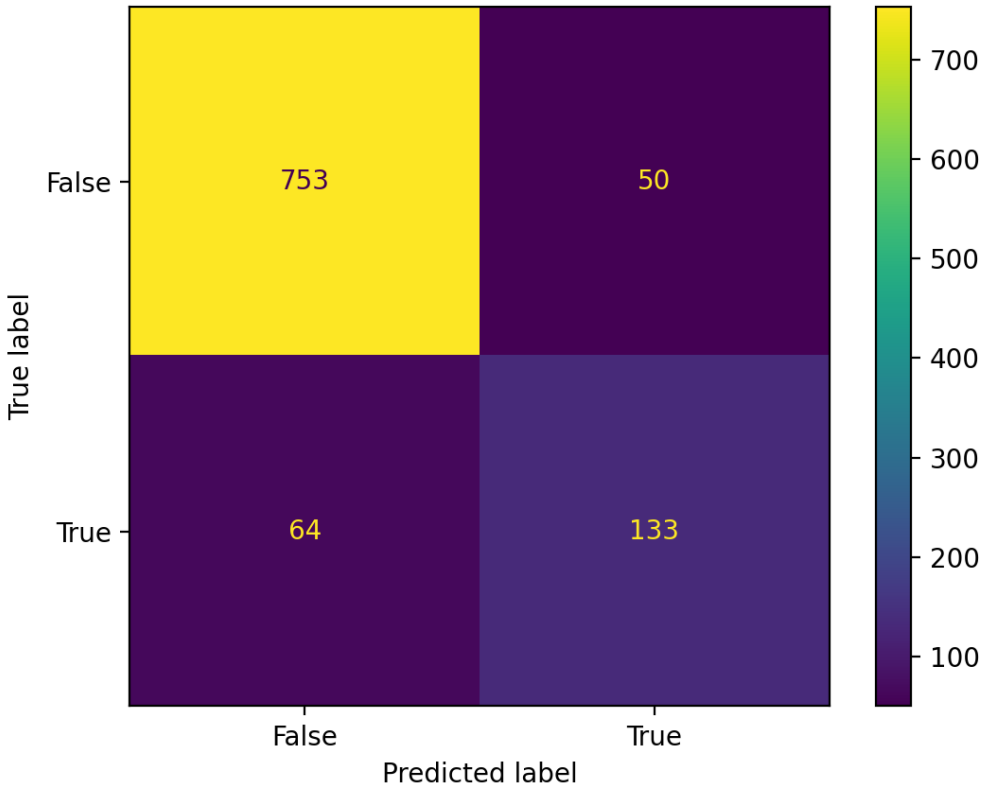
2. Wyniki

Wszystkie wyniki zbierane są jako statystyki dla testów przy uruchomieniu 25 razy. Dane były dzielone na zbiór treningowy i testowy w stosunku 4:1. Predykcja została wykonana na stworzonym na jednym z poprzednich ćwiczeń drzewie decyzyjnym ID3.

Tab. 2.1. Porównanie statystyk dla dokładność przy różnej ilości danych

Ilość danych	Średnia [%]	Std	Najlepszy	Najgorszy
50	86,0	11,31	100,0	50,0
100	85,4	7,06	95,00	70,00
200	87,0	5,20	97,50	77,50
300	86,34	4,16	93,33	78,33
500	86,44	2,73	92,00	80,00
1000	86,04	2,04	89,50	82,00
5000	86,33	1,05	88,40	84,40

Średnie wyniki są bardzo dobre niezależnie od ilości wygenerowanych danych - są one w okolicach 86% dokładności. Nie jest widoczny wzrost, ani spadek dokładności przy większej ilości danych. Zauważalny jest jednak spadek odchylenia standardowego w przypadku większych danych - dla tylko 50 próbek odchylenie to wynosi ponad 11, natomiast dla 5000 próbek zredukowało się ono do około 1. Również widoczny jest spadek maksymalnego wyniku, a wzrost minimalnego razem ze wzrostem ilości próbek.
W danych występowała duża przewaga klasy Ache z wartością False bo jest zgodne z tabelami prawdopodobieństwa. Dobre wyniki predykcji mogą wynikać ze sporych rozbieżności pomiędzy prawdopodobieństwami w tablicach prawdopodobieństwa np. szansa na to że Sport jest False to aż 0,98%, a wtedy prawdopodobieństwa na to że Back wyniesie False to 0,8 dla Chair równego True i 0,99 dla Char równego False. W związku z tym jest bardzo duże prawdopodobieństwo aby dane posiadały rekordy z wartością False dla Back, a wtedy szansa na uzyskanie wartości False dla rozpatrywanej klasy to również wysokie 90%. Na tej podstawie można stwierdzić, że dane nie są ciężkie do predykcji, a zwłaszcza przez drzewo decyzyjne, które ma nieograniczoną głębokość.



Rys. 2.1.

Na podstawie macierzy pomyłek widoczna jest wspominana dysproporcja pomiędzy wartościami False i True dla klasy Ache. Dodatkowo widoczne jest związane z tym częstsze popełnianie błędu związanego z przypisywaniem wartości False dla rekordów, które w rzeczywistości przyjmowały wartość True, niż na odwrót. Na podstawie tablic prawdopodobieństwa widać, że w przypadku kiedy B przyjęło wartość True, to rozbieżności prawdopodobieństwa w tym wypadku są już mniejsze, bo A przyjmie wartość True z prawdopodobieństwem 0,7, a False z 0,3. Może mieć to związek ze stosunkowo wysoką liczbą predykcji fałszywie pozytywnych.