WSI, Ćwiczenie 7

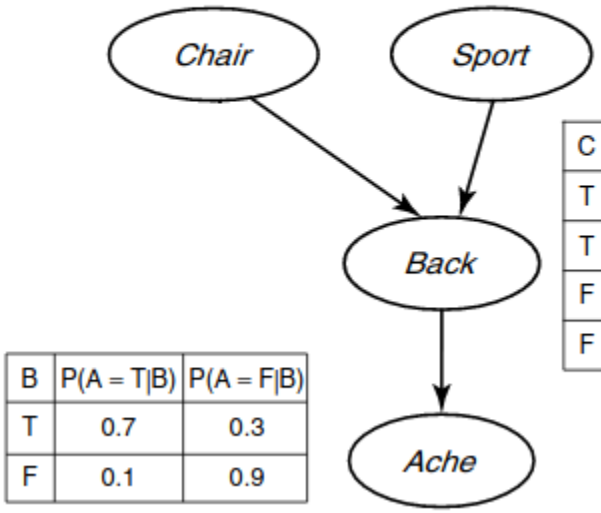
Michał Kwarciński

Polecenie:

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

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?

Wyniki:

Wyniki były zbierane dla testów uruchomionych 25 razy. Generowane dane były dzielone w stosunku 3:2 (40% testowych). Predykcja wykonana była na stworzonym w poprzednim ćwiczeniu drzewie decyzyjnym ID3.

Liczba iteracji	Średni wynik [%]	Najlepszy wynik [%]	Najgorszy wynik [%]	Odchylenie standardowe
25	81	97	55	10
50	85	89	70	4
100	84	91	74	4
250	84	89	80	2
500	85	88	82	2
1000	86	88	80	1
10000	86	86	85	0

Przykładowa macierz pomyłek dla danych testowych z uruchomienia dla 10000 próbek:

Prawdziwa wartość\Predykcja	True	False
True	530	225
False	365	2880

Wnioski:

Z tabeli można zauważyć, że średni wynik jest praktycznie niezależny od ilości wygenerowanych danych. Nie jest widoczny ani wzrost, ani spadek dokładności przy większej ilości danych (oprócz wyniku dla 25 iteracji – najprawdopodobniej jest to za mało danych do nauczenia drzewa). Zauważalny jest spadek odchylenia standardowego w przypadku większych danych - dla 25 próbek odchylenie to wynosi 10, natomiast dla 10000 próbek zredukowało się ono do prawie 0. Widoczny jest również spadek maksymalnego wyniku, a wzrost minimalnego wraz ze wzrostem liczby próbek.

Wyniki danych pokrywają się z zależnościami zawartymi w grafie. W wygenerowanych danych występowała duża przewaga klasy „Ache" z wartością False (około 80%). Prawdopodobieństwo, że „Sport" jest False to aż 98%. Wtedy prawdopodobieństwa na to, że Back będzie miało wartość False wynosi 80% dla Chair równego True (20%) i 99% dla Char równego False (80%). W związku z tym istnieje bardzo duże prawdopodobieństwo, że wartość „Back" była False. Wtedy szansa na uzyskanie wartości False dla rozpatrywanej klasy jest również wysoka (90%). Można stwierdzić, że te dane nie są trudne do predykcji.