WSI laboratorium 7 – Sprawozdanie

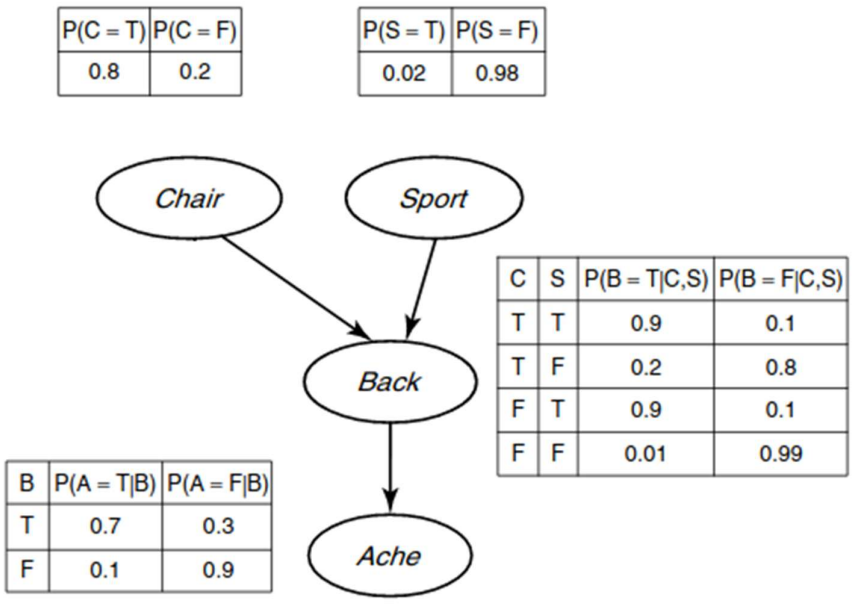
Pryimak Andrii-Stepan 336173

Wprowadzenie

Celem było stworzenie generatora, który działa na podstawie rozkładu reprezentowanego przez sieć bayesowską. Sieć ta opisuje zależności między zmiennymi losowymi w formie grafu, gdzie węzły sieci odpowiadają zmiennym, a krawędzie między nimi opisują zależności probabilistyczne.

Celem zadania było wygenerowanie zbioru danych, który będzie następnie użyty do treningu i testowania klasyfikatora z poprzednich ćwiczeń, wykorzystując zaimplementowany algorytm ID3.

Opis sieci Bayesa



Generator danych

Plik wejściowy (probabilities.json)

```
[
	{
		"id": "Chair",
		"parents": [],
		"probabilities": { "true": 0.8, "false": 0.2}
	},
	{
		"id": "Sport",
		"parents": [],
		"probabilities": { "true": 0.02, "false": 0.98}
	},
	{
		"id": "Back",
		"parents": ["Chair", "Sport"],
		"probabilities": {
			"TT": {"true": 0.9, "false": 0.1},
			"TF": {"true": 0.2, "false": 0.8},
			"FT": {"true": 0.1, "false": 0.9},
			"FF": {"true": 0.01, "false": 0.99}
		}
	},
	{
		"id": "Ache",
		"parents": ["Back"],
		"probabilities": {
			"T": {"true": 0.7, "false": 0.3},
			"F": {"true": 0.1, "false": 0.9}
		},
		"class": true
	}
]
```

Plik wyjściowy(generated_data.csv)

```
Chair,Sport,Back,Ache
False,False,False,False
True,False,False,False
...
```

Wygenerowane dane

Dystrybucja elementów dla 10 tys. Wygenerowanych linijek

Chair	Sport	Back	Ache	Rzeczywista	Oczekiwana ilość
True	True	True	True	98	100.8
True	True	True	False	41	43.2
True	True	False	True	1	1.6
True	True	False	False	13	14.4
True	False	True	True	1115	1097.6
True	False	True	False	487	470.4
True	False	False	True	671	627.2
True	False	False	False	5540	5644.8
False	True	True	True	3	2.8
False	True	True	False	0	1.2
False	True	False	True	5	3.6
False	True	False	False	33	32.4
False	False	True	True	18	13.72
False	False	True	False	6	5.88
False	False	False	True	220	194.04
False	False	False	False	1749	1746.36

Trening klasyfikatora

Liczba danych	Minimum	Średnia	Maximum	Odchylenia
10	0.75	0.81	1.0	0.1
50	0.65	0.87	1.0	0.06
100	0.75	0.85	0.95	0.04
500	0.82	0.86	0.91	0.017
1000	0.812	0.854	0.887	0.013
5000	0.844	0.860	0.877	0.006
10000	0.85	0.861	0.876	0.004
50000	0.861	0.866	0.871	0.001

Macierz pomyłek dla 50tys.

Predicted	False	True
Actual		
False	14868.03	1062.80
True	1604.51	2464.66

Warunkowa niezależność

1. Chair i Sport są niezależne (nie mają wspólnych rodziców, ani innych wspólnych zależności). (Chair⊥Sport)
2. Ache i Chair są warunkowo niezależne, biorąc pod uwagę Back. (Ache⊥Chair|Back)
3. Ache i Sport są warunkowo niezależne, biorąc pod uwagę Back. (Ache⊥Sport|Back)

Macierz pomyłek dla 50tys bez atrybutów Chair i Sport

Predicted	False	True
Actual		
False	14868	1062
True	1604	2464

Po usunięciu atrybutów Chair i Sport z procesu klasyfikacji wyniki w macierzy pomyłek nie zmieniają się znacząco, co potwierdza, że Ache jest warunkowo niezależne od Chair i Sport, przy znanym Back.

Podsumowanie

Celem laboratorium było stworzenie generatora danych działającego na podstawie sieci bayesowskiej, który generuje dane do treningu i testowania klasyfikatora opartego na algorytmie ID3. Zbudowana sieć bayesowska opisuje zależności między zmiennymi losowymi, takimi jak Chair, Sport, Back i Ache. Na podstawie tej sieci wygenerowano dane, które posłużyły do nauki klasyfikatora i testowania jego skuteczności.

Przeprowadzone testy klasyfikacyjne wykazały, że zbiory danych o różnych rozmiarach wpływają na skuteczność klasyfikatora. Dla mniejszych zestawów danych, klasyfikator osiągał maksymalną skuteczność na poziomie 100%, jednak wynikało to z losowości, ponieważ przy dużej liczbie powtórzeń występowały duże odchylenia standardowe. W miarę zwiększania rozmiaru zbioru danych skuteczność klasyfikatora stabilizowała się na poziomie 86,6%, a odchylenia standardowe malały.

Dodatkowo, analiza warunkowej niezależności wykazała, że zmienna Ache jest warunkowo niezależna od zmiennych Chair i Sport, gdy zmienna Back jest znana. Wykonano również analizę macierzy pomyłek po usunięciu kolumn warunkowo niezależnych. Wyniki były bardzo podobne do wcześniejszych, ponieważ, gdyby te zmienne były zależne, usunięcie jednej z nich spowodowałoby wzrost liczby błędów klasyfikacji, ponieważ klasyfikator nie miałby możliwości wykrywania zależności między tymi zmiennymi i wnioskowania drugiej na podstawie pierwszej.

Wyniki z laboratorium potwierdzają, że poprawnie zaimplementowany generator danych odzwierciedla zależności probabilistyczne w sieci bayesowskiej, a uzyskane dane mogą być skutecznie wykorzystywane do treningu klasyfikatora.