Studia podyplomowe Data Science
Sieci bayesowskie, odkrywanie przyczynowości

Na zajęciach poćwiczymy tworzenie sieci bayesowskich na podstawie wiedzy eksperta (włączywszy w to wiedzę z różnych źródeł, takich jak literatura fachowa) oraz uczenie sieci bayesowskich z danych/odkrywanie przyczynowości. Zadań jest więcej niż potrafimy wykonać, a więc proszę się nie przejmować tym, że Państwo nie dobrną do końca. Resztę proponuję wykonać w domu ☺.

Zadanie 1 (tworzenie sieci bayesowskich)
Policja może przeprowadzić proste badania zatrzymanego kierowcy w sytuacji podejrzenia o nadmierne spożycie alkoholu. Używa przy tym trzech metod: alkomatu, oceny oddechu i krótkiego spaceru po linii prostej. Alkomat jest najbardziej dokładny i wykrywa 90% nietrzeźwych kierowców. U 3% trzeźwych kierowców alkomat pokazuje jednak niesłusznie wynik pozytywny. Chuchnięcie jest mniej precyzyjne i mimo, że wykrywa pijanych kierowców z prawdopodobieństwem 80%, często, albowiem w 20% przypadków wykazuje ono wynik pozytywny u trzeźwych kierowców. Po linii prostej potrafi przejść 60% pijanych kierowcow i 90% trzeźwych kierowców. Statystyki mówią, że w dowolnym momencie na drogach jest około 0.5% nietrzeźwych kierowców.

Zbuduj model sieci bayesowskiej, który pomoże w odpowiedzi na następujące pięć pytań:

(a) Pan Zdzisław poddany został badaniu alkomatem, który wykazał wynik pozytywny. Jednak w jego oddechu policjant nie poczuł alkoholu. Jakie jest prawdopodobieństwo tego, że Pan Zdzisław jest nietrzeźwy?

(b) Ponieważ policjant miał wrażenie, że Pan Zdzisław był wstawiony, nalegał na przejście po linii prostej, co Pan Zdzisław zrobił bez problemu. Jakie jest teraz prawdopodobieństwo tego, że Pan Zdzisław jest nietrzeźwy?

(c) Pani Prudencja poddana była również testowi alkomatem, który wykazał wynik negatywny. Policjant jednak poczuł alkoholu w jej oddechu. Potrafiła ona również przejść po linii prostej. Jakie jest prawdopodobieństwo tego, że jest ona nietrzeźwa?

(d) W wypadku Pana Felka, wszystkie trzy proste testy, tzn. alkomat, oddech, jak i przejście po linii prostej wykazały wyniki pozytywne. Jakie jest prawdopodobieństwo tego, że Pan Felek jest nietrzeźwy?

(e) W wypadku Pana Donalda, wszystkie trzy testy pokazały wynik negatywny. Jakie jest prawdopodobieństwo tego, że Pan Donald jest nietrzeźwy?

Zadanie 2 (odkrywanie przyczynowości)
Załączony plik `gry.txt` zawiera zbiór danych zebranych wśród studentów Politechniki Białostockiej. Plik zawiera następujące zmienne:

`Zamieszkanie`: Miejsce zamieszkania (`Miasto/Wies`)
`Komputer`: Czy posiada własny komputer (`Ma/Nie`)
`Zamoznosc`: Czy pochodzi z zamożnej rodziny (`Bogaty/Biedny`)
`CzasRodzicow`: Czas, jaki studentowi poswięcają rodzice (`Duzo/Malo`)
`PelnaRodzina`: Czy są w pobliżu oboje ojciec i matka (`Tak/Nie`)
`Marginalizowanie`: Czy student jest marginalizowany w domu (`Rodzenstwo/Rodzice/Nie`)
`KontaktZRodzina`: Czy ma dobry kontakt z rodziną (`Zly/Dobry`)
`Uzaleznienie`: Czy jest uzależniony od gier komputerowych (`Tak/Nie`)
`Czas`: Czas spędzony przed komputerem (`Duzo/Malo`)
`Oceny`: Oceny na studiach (`Dobre/Slabe`)
`Kondycja`: Kondycja fizyczna (`Dobra/Slaba`)
`Wzrok`: Stan wzroku (`Dobry/Slaby`)
`Przyjaciele`: Czy posiada przyjaciół (`Ma/NieMa`)
`Klopoty`: Czy ma kłopoty osobiste (`Koledzy/Szkola/Brak`)

`Sen`: Ilość snu (`PiecMinus/Szesc/SiedemPlus`)
`Postawa`: Czy student ma wady postawy (`Paragraf/Prosta`)

Odpowiedz na następujące pytania:

(a) Które ze zmiennych wydają się być bezpośrednimi przyczynami uzależnienia od gier komputerowych?
(b) Które ze zmiennych są najlepszymi źródłami informacji w przewidywaniu uzależnienia od gier komputerowych?
(c) Z jaką dokładnością można przewidzieć czy student jest uzależniony od gier komputerowych (podaj ogólną dokładność przewidywania, czułość oraz specyficzność)?

Zadanie 3 (uczenie sieci bayesowskich z danych, walidacja krzyżowa)
Załączony plik `house-votes.txt` zawiera zbiór danych z repozytorium `Machine Learning Repository` (`University of California, Irvine`), opisujacy głosy oddane przez kongresmenów amerykańskich w 16 różnych sprawach. Utwórz sieć bayesowską, która będzie klasyfikować dane względem partii, do której należał kongresman (zmienna `class`), na podstawie jego głosów oddanych w sprawach uwzględnionych w bazie danych.

Przetestuj otrzymane wyniki używając wbudowanej funkcji oceny modelu na podstawie danych (*Validate* z menu *Data*). Która z metod uczenia pozwala na osiągnięcie najwyższej dokładności w klasyfikacji?

Wskazówki do Zadania 1: Ponieważ algorytmy do uczenia struktury sieci w **GeNIe** (jeszcze ☺) nie są w stanie uczyć się ze zbiorów z danymi brakującymi, dane brakujące mogą zostać potraktowane w następujący sposób:
(1) usuń rekordy z danymi brakującymi,
(2) naucz się struktury sieci,
(3) załaduj ponownie zbiór danych z danymi brakującymi i użyj tego zbioru do wyznaczenia parametrów istniejącego modelu.

Zadanie 4 (odkrywanie przyczynowości)
Na podstawie pliku danych `retention.txt` sprawdź, które ze zmiennych zawartych w pliku są przyczynami rezygnacji ze studiów studentów na uczelniach amerykańskich.

Znaczenie zmiennych w pliku `retention.txt`:
`spend`: całkowita suma w $USD wydawana średnio na jednego studenta w ciągu roku na uczelni (miara zasobów uczelni)
`apret`: procent studentów, którzy kończą pierwszy rok studiów i pozostają na uczelni
`top10`: procent studentów pierwszego roku, którzy byli w 10% najlepszych studentów w swojej szkole średniej (miara jakości studentów)
`rejr`: procent kandydatów na studia którzy nie zostali przyjęci na studia (miara selektywności uczelni: im wyższy procent, tym bardziej selektywna uczelnia)
`tstsc`: wyniki standardowych testów dla kandydatów na studia, wyrażone w skali 0-100 (100 to wynik najlepszy; miara jakości studentów)
`pacc`: procent kandydatów przyjętych na studia przez uczelnię, którzy akceptują ofertę uczelni i rozpoczynają studia (na uczelniach amerykańskich studenci składają typowo podania na kilka uczelni i akceptują najlepszą oferte; miara prestiżu uczelni)
`strat`: stosunek liczby studentów do liczby wykładowców (liczba studentów przypadajacą na jednego wykładowcę; miara jakości dydaktyki)
`salar`: średnia płaca wykładowcy na uczelni w $USD (miara jakości kadry)