WSTĘP DO SZTUCZNEJ INTELIGENCJI

Ćwiczenie7 – Sieć Bayesowska

JAKUB KWAŚNIAK 331396

Wstęp

Treść zadania:

Dla zbioru danych o zabójstwach w USA z lat 1980 – 2014 https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset wybrać następujące cechy {Victim Sex, Victim Age, Victim Race, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Relationship, Weapon}

Przy pomocy jednej z bibliotek pgmpy, pomegranate, bnlearn wygenerować sieć Bayesowską modelującą zależności pomiędzy tymi cechami. Podpowiedź: należy znaleźć strukturę sieci (structure learning), następnie estymować prawdopodobieństwa warunkowe pomiędzy zmiennymi losowymi (parameter learning).

Zwizualizować i przeanalizować nauczoną sieć - jakie są rozkłady prawdopodobieństw pojedynczych cech, jakie zależności pomiędzy cechami można zauważyć?

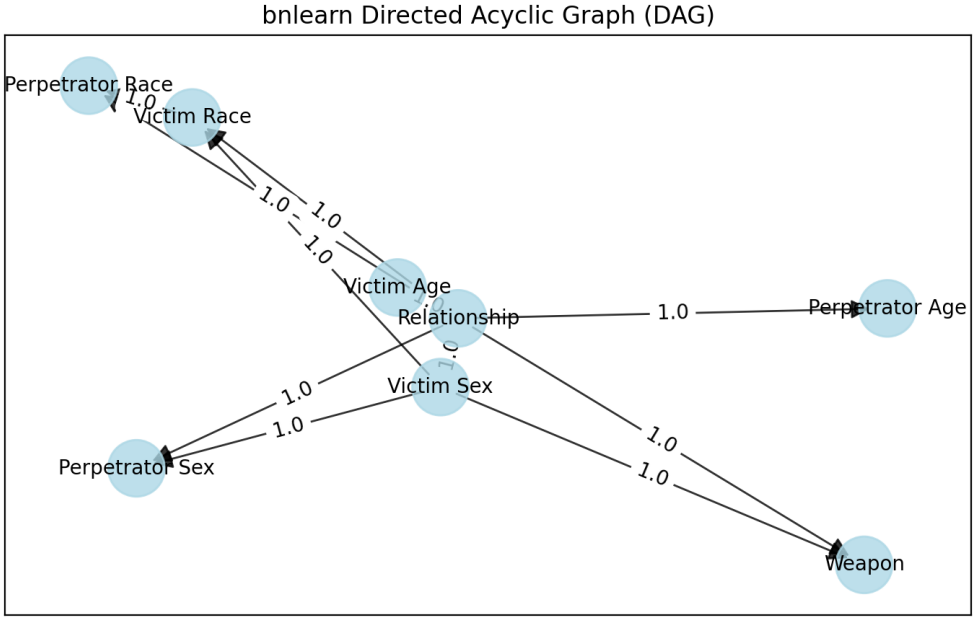
Zaimplementować losowy generator danych, który działa zgodnie z rozkładem reprezentowanym przez wygenerowaną sieć.

Użyć generatora do wygenerowania kilku losowych morderstw, podając jako argumenty różne obserwacje.

Wyniki

Eksperyment dla całej bazy danych:

Wygenerowany graf skierowany – znaleziona struktura sieci (połączenia między węzłami – pokazują które cechy/zmienne losowe są zależne od innych), oraz wagi określające prawdopodobieństwo warunkowe pomiędzy zmiennymi losowymi (wartości na krawędziach – wskazują również kierunek zależności, która cecha jest rodzicem, a która dzieckiem)



```
Network structure: target           Victim Sex  Victim Age  ...  Relationship  Weapon
source                                        ...
Victim Sex             False       False  ...         False    True
Victim Age             False       False  ...          True   False
Victim Race            False       False  ...         False   False
Perpetrator Sex        False       False  ...         False   False
Perpetrator Age        False       False  ...         False   False
Perpetrator Race       False       False  ...         False   False
Relationship            True       False  ...         False    True
Weapon                 False       False  ...         False   False
```

Widoczna powyżej jest nauczona struktura sieci gdzie w kolumnie source widoczne są zmienne losowe a nagłówki pozostałych kolumn to targety (zmienne warunkowe). W wierszach zdefiniowane są zależności – jeśli True zmienna losowa z target zależy od zmiennej losowej z source, jeśli False nie zależy

Jak widać dla cechy np. Victim Sex – jest ona zależna od Relashionship, z kolei Weapon zależne jest od Victim Sex oraz Relationship co również potwierdza graf, fragmenty tabel z estymowanym prawdopodobieństwem warunkowym zostały spotokowane do pliku „results_full_db.txt”

Oto 3 otrzymane wyniki dla podanych niepełnych obserwacji, losowane z rozkładem prawdopodobieństwa sieci Bayesa:

```
Niepełne obserwacje: {'Victim Sex': 'Male', 'Perpetrator Age': 30, 'Relationship': 'Wife'}
Otrzymane przewidywania: Victim Age               34
Victim Race            White
Perpetrator Sex         Male
Perpetrator Race       White
Weapon              Shotgun
p                  0.002509
Name: 16797, dtype: object
Niepełne obserwacje: {'Victim Age': 20, 'Perpetrator Race': 'Black', 'Weapon': 'Knife'}
Otrzymane przewidywania: Victim Sex               Female
Victim Race               Black
Perpetrator Sex            Male
Perpetrator Age              39
Relationship       Acquaintance
p                      0.000977
Name: 9153, dtype: object
Niepełne obserwacje: {'Victim Age': 50, 'Victim Sex': 'Female', 'Perpetrator Race': 'White'}
Otrzymane przewidywania: Victim Race           White
Perpetrator Sex        Male
Perpetrator Age          69
Relationship         Mother
Weapon              Handgun
p                  0.000013
Name: 311708, dtype: object
```

Obserwacje i wnioski

1. Największe prawdopodobieństwo morderstwa zaobserwować można w obrębie jednej rasy
2. Rozkład cech „Perpetrator Age” oraz „Victim Age” są rozkładami dyskretnymi o szerokim przedziale wartości przez to mogą pojawić się problemy przy generowaniu krotki z niepełnej obserwacji – sieć „zamarza” i nie jest w stanie wygenerować ‘query’ z metodą bn.inference.fit() – nie może określić z jakim prawdopodobieństwem rozłożą się pozostałe cechy dla podanych obserwacji
   W celu zapobiegnięcia temu należy do każdej obserwacji podawać przynajmniej jednną zmienną losową „… Age” lub zawęzić rozkład dyskretny zmiennych „… Age” np. filtrując dataframe tylko do wartości „… Age” < 30 lub zawężając ilość danych do np. 1/3 wierszy dataframe’u
3. Z powodu opisanego wyżej z modelu wygenerowanego przez sieć pracująca na pełnym zbiorze danych nie będziemy w stanie wygenerować krotki dla pustej obserwacji.

- Sieć licząc prawdopodobieństwo warunkowe np. płci ofiary (Victim Sex) pod warunkiem związku z mordercą (Relationship – przy czym Relationship(Wife) oznacza, że ofiara była żoną mordercy) przydziela niskie ale nie zerowe prawdopodobieństwo sytuacjom niemożliwym – np. P (Victim_Sex(Female) | Relationship(Wife)) = 0.972633781 ale jednocześnie P (Victim_Sex(Male) | Relationship(Wife)) = 0.027366219.
  Może to wynikać z faktu że między zależnymi od siebie zmiennymi losowymi sieć przeprowadza permutacje wszystkich możliwych wartości tworząc też scenariusze niemożliwe i przypisuje im marginalne ale nie zerowe prawdopodobieństwo, żeby nie dyskwalifikować kombinacji potencjalnie możliwej w przestrzeni probabilistycznej – w końcu sieć nie ma intuicji, że żona płci męskiej nie istnieje (chociaż to zbór danych z USA to kto wie), nie ma takiej kombinacji w danych ale sama kombinacja zmiennych bez dodatkowych warunków mogłaby się pojawić