

WSTĘP DO SZTUCZNEJ INTELIGENCJI

Ćwiczenie 7 – 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?

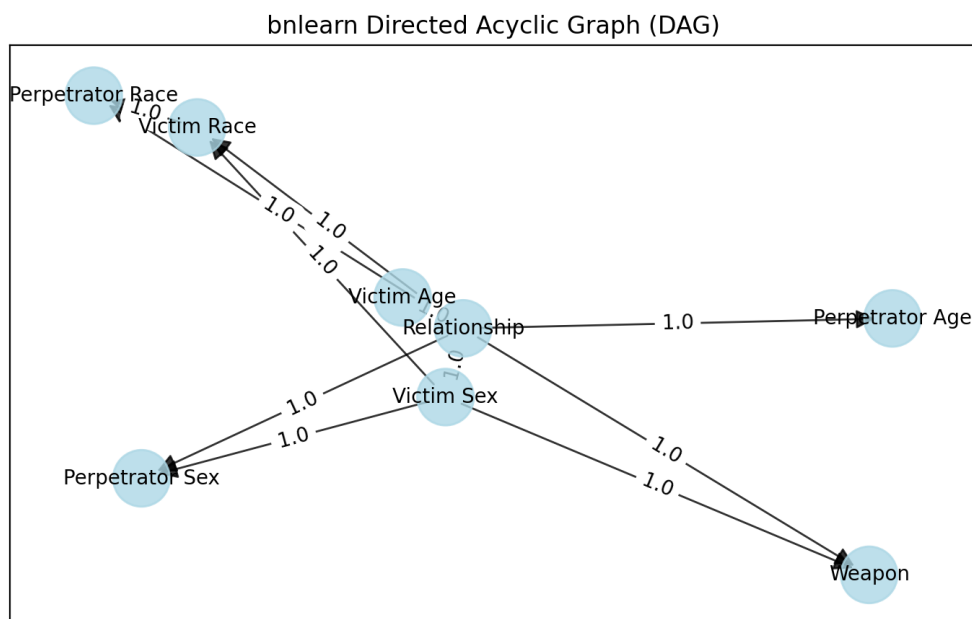
Zaimplementować losowy generator danych, który działa zgodnie z rozkładem reprezentowanym przez wygenerowaną sieć.

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)



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 jedną 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 pracującą 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. $P(\text{Victim_Sex(Female)} | \text{Relationship(Wife)}) = 0.972633781$ ale jednocześnie $P(\text{Victim_Sex(Male)} | \text{Relationship(Wife)}) = 0.027366219$ (sytuacja teoretycznie niemożliwa, chociaż jest to zbiór danych z USA, więc pewnie i dziwniejsze relacje mogłyby się pojawić)