WSTĘP DO SZTUCZNEJ INTELIGENCJI

Ćwiczenie 7 – Sieć Bayesowska

MICHAŁ MIZIA 331407

SPIS TREŚCI

Wstęp

1. 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}
2. 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).
3. Zwizualizować i przeanalizować nauczoną sieć - jakie są rozkłady prawdopodobieństw pojedynczych cech, jakie zależności pomiędzy cechami można zauważyć?
4. Zaimplementować losowy generator danych, który działa zgodnie z rozkładem reprezentowanym przez wygenerowaną sieć.
5. Użyć generatora do wygenerowania kilku losowych morderstw, podając jako argumenty różne obserwacje.

IMPLEMENTACJA SIECI

Zbiór danych morderstw z USA zawiera niektóre dane niepełne oraz niepoprawne, ze względu na to należy go oczyścić przed rozpoczęciem analizy. Pobrany plik jest czyszczony jednokrotnie w pliku parse_data.py a następnie zapisywany jako nowy plik który następnie służy do trenowania sieci.

- parse_data.py

```
df = df[
    [
        "Victim Sex",
        "Victim Age",
        "Victim Race",
        "Victim Ethnicity",
        "Perpetrator Sex",
        "Perpetrator Age",
        "Perpetrator Race",
        "Perpetrator Ethnicity",
        "Relationship",
        "Weapon",
    ]
]
df = df[
    ~df.apply(
        lambda row: row.astype(str).str.contains("unknown", case=False).any(),
        axis=1,
    )
]
df = df[(df["Victim Age"] != 0) & (df["Perpetrator Age"] != 0)]
df.to_csv("data/US_Crime_Data.csv", index=False)
```

Zachowujemy wybrane kolumny, usuwamy wiersze z wartościami unknown oraz wiersze w których wiek wynosi 0.

Do trenowania sieci użyłem biblioteki **bnlearn.**

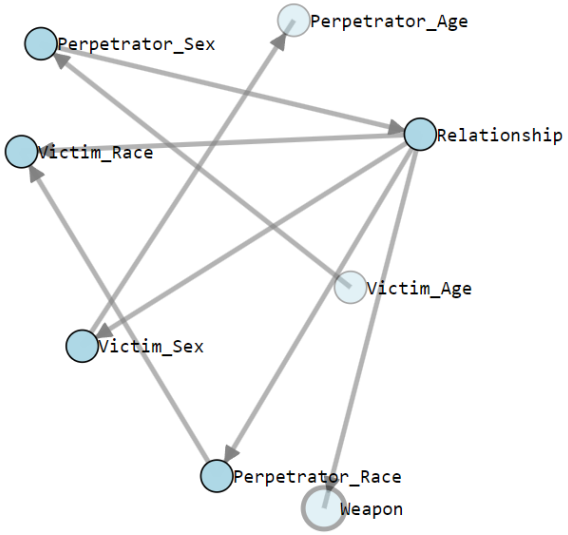
W pierwszej wersji rozwiązania przypadkowo zostawiłem parametry perpetrator/victim ethnicity. Doprowadziło to do ciekawej obserwacji, kiedy liczba kombinacji w cpd jest zbyt duża, sieć nie była w stanie stworzyć powiązań między parametrami i omijała niektóre z nich. W związku z tym próbowałem

zdyskretyzować wiek do przedziałów 5 letnich, np. 10-15 itd. Przy tym rozwiązaniu model był już w stanie nauczyć się wszystkich kombinacji i estymować wiek.

W ostatecznej wersji perpetrator/victim ethicity są usunięte a wiek nie jest bardziej dyskretyzowany ale uważam to za ciekawą obserwację.

Wyniki

Wygenerowana sieć:



Przykładowe wygenerowane dane (na zielono oznaczone są dowody):

Victim Age	Perpetrator Age	Relationship	Perpetrator Sex	Victim Race	Weapon	Perpetrator Race	Victim Sex
20	20	Husband	Female	Black	Knife	Black	Male
20	20	Husband	Female	White	Handgun	White	Male
20	20	Husband	Female	Black	Fire	Black	Male
20	20	Husband	Female	White	Handgun	White	Male
20	20	Husband	Female	Black	Handgun	Black	Male

Victim Race	Perpetrator Age	Weapon	Victim Sex	Victim Age	Perpetrator Race	Perpetrator Sex	Relationship
Black	20	Knife	Female	46	Black	Male	Wife
Black	20	Knife	Male	43	Black	Male	Acquaintance
Black	20	Knife	Male	42	Black	Female	Acquaintance
Black	20	Knife	Male	43	Black	Male	Boyfriend
Black	20	Knife	Male	30	Black	Male	Acquaintance

Victim Sex	Perpetrator Age	Weapon	Victim Race	Perpetrator Race	Perpetrator Sex	Victim Age	Relationship
Male	20	Shotgun	White	White	Male	21	Stranger
Male	20	Shotgun	Black	Black	Male	20	Neighbor
Male	20	Shotgun	Black	Black	Male	15	Neighbor
Male	20	Shotgun	Black	Black	Male	85	Stranger
Male	20	Shotgun	White	White	Male	67	Friend

Wnioski

Kiedy sieć ma dużą ilość parametrów przy czym niektóre z nich są całkowitoliczbowe albo ciągłe, mocniejsze zdyskretyzowanie ich może mocno pomóc w trenowaniu sieci. Przy za dużej liczbie kombinacji sieć zaczyna gubić parametry.

Morderstwa dzieją się głównie wewnątrz kręgów rasowych, jeżeli ofiara to native american to mimo dużo większej ilości osób innych ras, największe pradopodobieństwo na rasę sprawcy ma native american.

Wartości wieku są słabo powiązane ze względu na bardzo dużą liczbę przyjmowanych wartości.

Najwięcej morderstw jest dokonywanych na osobach obcych.