Sprawozdanie ćwiczenia nr 7 – Łukasz Szydlik

Cel i opis eksperymentów

Celem ćwiczenia jest:

1. Dla zbioru danych o zabójstwach w USA z lat 1980-2014 wybrać następujące cechy {Victim Sex, Victim Age, Victim Race, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Relationship, Weapon} oraz sprawdzić wiarygodność danych i odrzucić ewentualne błędy.

   https://www.kaggle.com/datasets/mrayushagrawal/us-crime-dataset

2. Przy pomocy jednej z bibliotek pgmpy, pomegranate, bnlearn wygenerować sieć Bayesowską modelującą zależności pomiędzy tymi cechami.
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.

Instrukcja programu

Instalacja

1. Należy pobrać repozytorium

2. Pobrać plik danych oraz umieścić go w folderze lab7 pod nazwą `US_Crime_DataSet.csv`

3. Upewnić się, że posiadamy wymagane biblioteki:

`pip install -r requirements.txt`

4. Zainstalować rozszerzenie Jupyter (VS Code)

Uruchomienie programu

W celu uruchomienia BayesianNetwork.ipynb wymagane jest korzystanie z notatnika Jupyter

W Jupyter ustawiamy kernel na nasze środowisko python. Następnie po naciśnięciu ”Run All” będziemy mogli zauważyć wygenerowaną sięć (Visualization of network) oraz rozkłady prawdopodobieństw warunkowych (CPDs of variables).

W ostatniej sekcji `Data Generator` jest pokazany przykład generowania przykładowych danych zgodnie rozkładami prawdopodobieństw naszej sieci. Możemy również umieścić naszą obserwację w pliku `observation.json`. Ustawiamy ”?” w danych, które chcemy przewidzieć. Następnie możemy uruchomić ostatnią komórkę w notatniku, aby otrzymać wygenerowany przykład w pliku `generated_sample.json`.

Wyniki ćwiczenia

Rozkłady prawdopodobieństw:
(liczby po `:` w prawdopodobieństwie oznaczają ilość unikalnych cech):

P(Victim Sex:2 | Relationship:27, Victim Age:100)

	Victim Sex	Relationship	Victim Age	Probability
0	Female	Acquaintance	0	0.45685279187817257
1	Female	Acquaintance	1	0.4576923076923077
2	Female	Acquaintance	2	0.4742489270386266
3	Female	Acquaintance	3	0.35826771653543305
4	Female	Acquaintance	4	0.3652694610778443
...	...	...	...	...

P(Victim Age:100)

	Victim Age	Probability
0	0	0.015967
1	1	0.009662
2	2	0.007154
3	3	0.004582
4	4	0.003360
...	...	...

P(Victim Race:4 | Relationship:27, Weapon:15)

	Victim Race	Relationship	Weapon	Probability
0	Asian/Pacific Islander	Acquaintance	Blunt Object	0.011945522331279008
1	Asian/Pacific Islander	Acquaintance	Drowning	0.015748031496062992
2	Asian/Pacific Islander	Acquaintance	Drugs	0.0
3	Asian/Pacific Islander	Acquaintance	Explosives	0.0
4	Asian/Pacific Islander	Acquaintance	Fall	0.0
...	...	...	...	...

P(Perpetrator Sex:2 | Perpetrator Race:4, Relationship:27, Victim Race:4, Weapon:15)

	Perpetrator Race	Relationship	Victim Race	Weapon	Perpetrator Sex	Probability
0	Asian/Pacific Islander	Acquaintance	Asian/Pacific Islander	Blunt Object	Female	0.12345679012345678
1	Asian/Pacific Islander	Acquaintance	Asian/Pacific Islander	Drowning	Female	0.0
2	Asian/Pacific Islander	Acquaintance	Asian/Pacific Islander	Drugs	Female	0.5
3	Asian/Pacific Islander	Acquaintance	Asian/Pacific Islander	Explosives	Female	0.5
4	Asian/Pacific Islander	Acquaintance	Asian/Pacific Islander	Fall	Female	0.5
...	...	...	...	...	...	...

P(Perpetrator Age:99)

	Perpetrator Age	Probability
0	1	0.000012
1	2	0.000018
2	3	0.000068
3	4	0.000080
4	5	0.000086
...	...	...

P(Perpetrator Race:4 | Relationship:27, Victim Race:4)

	Perpetrator Race	Relationship	Victim Race	Probability
0	Asian/Pacific Islander	Acquaintance	Asian/Pacific Islander	0.6652433817250214
1	Asian/Pacific Islander	Acquaintance	Black	0.0015589730265187808
2	Asian/Pacific Islander	Acquaintance	Native American/Alaska Native	0.005454545454545455
3	Asian/Pacific Islander	Acquaintance	White	0.005251441292311205
4	Asian/Pacific Islander	Boyfriend	Asian/Pacific Islander	0.675
...	...	...	...	...

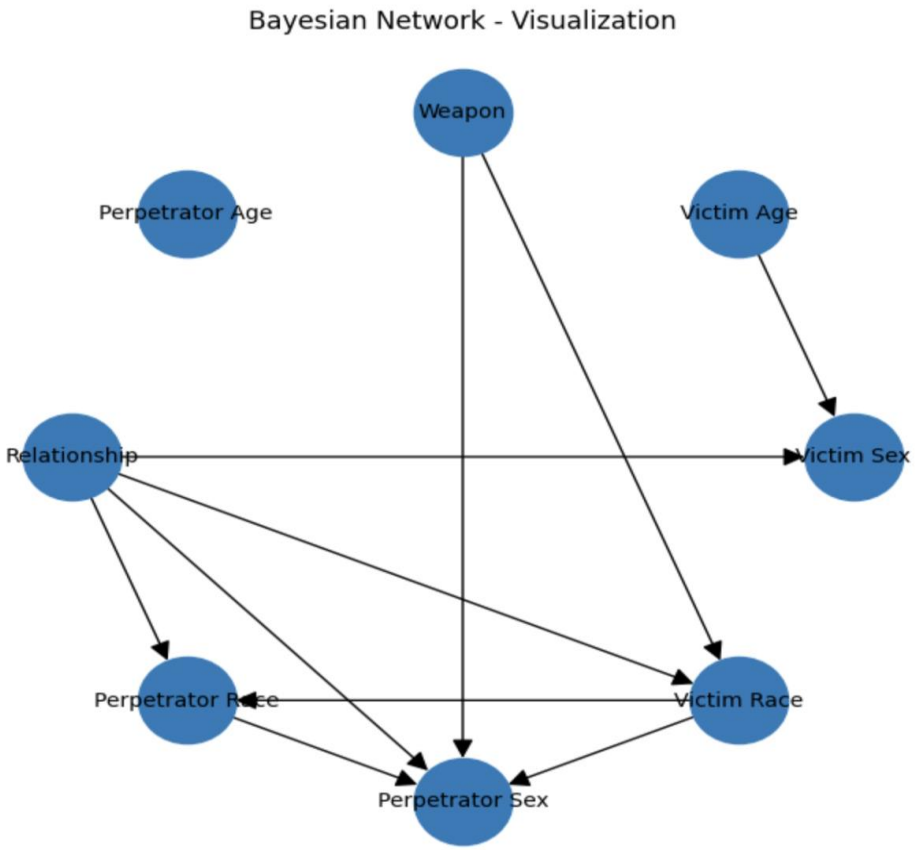
P(Relationship:27)

	Relationship	Probability
0	Acquaintance	0.358204
1	Boyfriend	0.021377
2	Boyfriend/Girlfriend	0.003705
3	Brother	0.016281
4	Common-Law Husband	0.005782

P(Weapon:15)

	Weapon	Probability
0	Blunt Object	0.125333
1	Drowning	0.002791
2	Drugs	0.003560
3	Explosives	0.001022
4	Fall	0.000446

Wizualizacja oraz analiza sieci:



1. Rozkłady prawdopodobieństw

Na podstawie przykładowych rozkładów możemy zauważyć, że są znormalizowane, czyli sumują się do 1. Natomiast niektóre nie prawdopodobieństwami warunkowymi i są obliczone jedynie na podstawie ich występowania w bazie danych.

2. Zależności

Możemy zauważyć, że `Perpetrator Age` nie ma żadnych połączeń w sieci co oznacza, że jest on zmienną niezależną od reszty. Węzły `Weapon`, `Victim Age` oraz `Relationship są warunkowo zależne, czyli niezależne gdy nie ma podanego ich dziecka, ponieważ są to węzły bez rodziców. Natomiast pozostałe węzły są zależne.

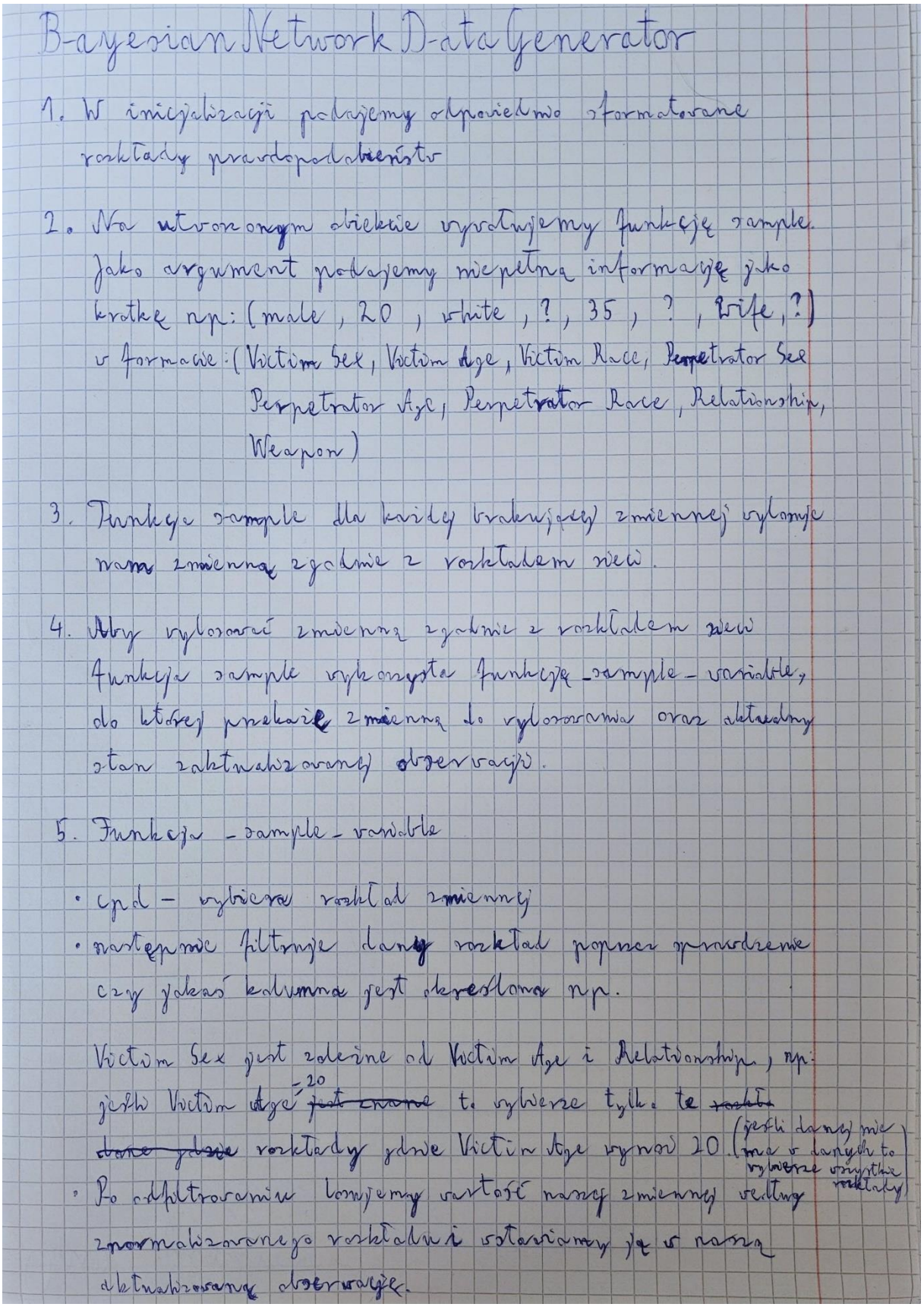
3. Wspólna przyczyna

Występuje wtedy gdy dwa węzły mają tego samego rodzica oraz nie są zależne względem siebie. Oznacza to, że dzieci mogą być warunkowo zależne, czyli niezależne jeśli nie znamy wartości rodzica. W naszej sieci ta sytuacja nie występuje. Są węzły które mają wspólnego rodzica, ale są zależne od siebie.

3. Wspólny skutek

Występuje, gdy dwa węzły są rodzicami wspólnego węzła oraz są niezależne od siebie. Oznacza to, że rodzice mogą być warunkowo zależni, czyli niezależne jeśli nie znamy wartości dziecka. W naszej sieci obrazują to węzły `Relationship`, Victim Age` oraz `Victim Sex`.

Generator danych:



Bayesian Network Data Generator

1. W inicjalizacji podajemy odpowiednio sformatowane rozkłady prawdopodobieństw

2. Na utworzonym obiekcie wywołujemy funkcję sample. Jako argument podajemy niepełną informację jako krotkę np: (male, 20, white, ?, 35, ?, Rifle, ?) w formacie: (Victim Sex, Victim Age, Victim Race, Perpetrator Sex, Perpetrator Age, Perpetrator Race, Relationship, Weapon)

3. Funkcja sample dla każdej brakującej zmiennej wylosuje nam zmienną zgodnie z rozkładem wież.

4. Aby wylosować zmienną zgodnie z rozkładem wież funkcja sample wykorzysta funkcję _sample_variable, do której przekaże zmienną do wylosowania oraz aktualny stan zaktualizowanej obserwacji.

5. Funkcja _sample_variable

- cpd – wybiera rozkład zmiennej
- następnie filtruje dany rozkład pomiar prawdziwie czy jakaś kolumna jest określona np.

  Victim Sex jest zależne od Victim Age i Relationship, np: jeśli Victim Age = 20 to wybierze tylko te rozkłady gdzie Victim Age wynosi 20. (jeśli danej nie ma w danych to wybierze wszystkie rozkłady)
- Po odfiltrowaniu losujemy wartość naszej zmiennej według znormalizowanego rozkładu i ustawiamy ją w naszą aktualizowaną obserwację.

Przykładowe obserwacje:

```
observation1 = ("?", "20", "?", "male", "?", "asian", "friend", "strangulation")
observation2 = ("Male", "?", "white", "?", "?", "black", "?", "handgun")
observation3 = ("Female", "42", "black", "?", "30", "?", "?", "?")
observation4 = ("White", "?", "Animal", "?", "10000", "?", "?", "knife")
observation5 = ("?", "?", "?", "?", "?", "?", "?", "?")
```

Wygenerowane przewidzenia:

```
Generated sample1: ('Female', '20', 'Asian/Pacific Islander', 'Male', '23', 'Asian', 'Friend', 'Strangulation')
Generated sample2: ('Male', '65', 'White', 'Asian/Pacific Islander', '38', 'Black', 'Husband', 'Handgun')
Generated sample3: ('Female', '42', 'Black', 'White', '30', 'Black', 'Acquaintance', 'Knife')
Generated sample4: ('White', '25', 'Animal', 'White', '10000', 'White', 'Acquaintance', 'Knife')
Generated sample5: ('Female', '84', 'White', 'Native American/Alaska Native', '23', 'White', 'Family', 'Blunt Object')
```

Wnioski

1) **Zależności:** Sieci Bayesowskie pozwalają na reprezentowanie i analizowanie relacji między zmiennymi. Każda krawędź w sieci reprezentuje potencjalny związek przyczynowo-skutkowy.
2) **Znaczenie jakości danych:** Dokładność struktury i parametrów sieci jest silnie uzależniona od jakości danych wejściowych. Brakujące dane lub błędy mogą znacząco wpłynąć na wyniki.
3) **Wpływ zmiennych ukrytych:** Jeśli niektóre zmienne są niewidoczne w danych (zmienne pominięte), mogą one prowadzić do nieprawidłowych relacji w modelu, co skutkuje powstawaniem zmiennych niezależnych.
4) **Uczenie struktury sieci:** Sieci Bayesowskie mogą być uczone z danych za pomocą zaawansowanych algorytmów, które pozwalają na odkrywanie nieoczywistych relacji między zmiennymi.
5) **Przewidywanie i uzupełnianie brakujących danych:** Sieci Bayesowskie umożliwiają wnioskowanie probabilistyczne, co pozwala na przewidywanie brakujących danych.
6) **Identyfikacja prawdopodobieństw warunkowych:** Dzięki inferencji możliwe jest obliczanie prawdopodobieństw warunkowych dla różnych scenariuszy, co jest szczególnie przydatne w diagnostyce, podejmowaniu decyzji i prognozowaniu.
7) **Możliwość włączania wiedzy eksperckiej:** Sieci Bayesowskie umożliwiają włączanie wiedzy eksperckiej do modelu w postaci własnych krawędzi.