Metody Inżynierii Wiedzy Modelowanie probabilistyczne, sieci Bayesa, łańcuchy markowa HMM - wykład 3

Adam Szmigielski
aszmigie@pjwstk.edu.pl
materiały: $ftp(public) : //aszmigie/MIW$

Niepewność

- Niepewność w rzeczywistych procesach decyzyjnych,
- Niepewność jako cecha informacji,
- Niepewność jako wada informacji,
- Jak wnioskować (podejmować decyzje) w warunkach niepewności ?

Prawdopodobieństwo

- Zdarzenie elementarne jako zdanie wartościowalne w logice dwuwartościowej,
- Prawdopodobieństwo może modelować sytuacje konfliktowe,
- Różne interpretacje prawdopodobieństwa
  1. subiektywne,
  2. częstotliwościowe,
  3. jako miara (formalna definicja)

Definicja prawdopodobieństwa

Miarą probabilistyczną $p : \varphi \mapsto [0,1]$ nazywamy funkcję określoną na podzbiorach n-elementowym zbioru X i spełniającą własności:

WP1: $p(\emptyset) = 0$,

WP2: $p(X) = 1$ - postulat unormowania,

WP3: Prawdopodobieństwo sumy przeliczalnej ilości zdarzeń parami rozłącznych równe jest sumie prawdopodobieństw tych zdarzeń tzn.

$$p(\cup_{i=0}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i).$$

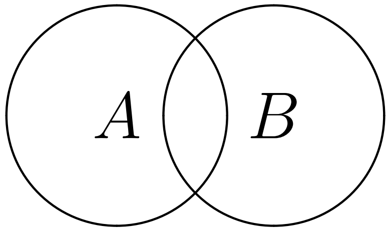
\- *postulat przeliczalnej addytywności.*

Prawdopodobieństwo apriori

- $p(A)$ jest prawdopodobieństwem apriori zdarzenia A (np. $p(A) = 0.1$),
- oznacza: Że szansa zajścia zdarzenia A wynosi 0.1,
- $p(A)$ może być użyte jeśli nie mamy żadnej, dodatkowej informacji.
- Np.: Dla zmiennej losowej *Pogoda* możemy mieć następujące prawdopodobieństwa apriori:
  p(Pogoda słoneczna) = 0.5
  p(Pogoda deszczowa) = 0.2
  p(Pogoda zachmurzona) = 0.1
  p(Pogoda Śnieżna) = 0.2

Prawdopodobieństwo warunkowe

- prawdopodobieństwo A jeśli wiemy, że B.



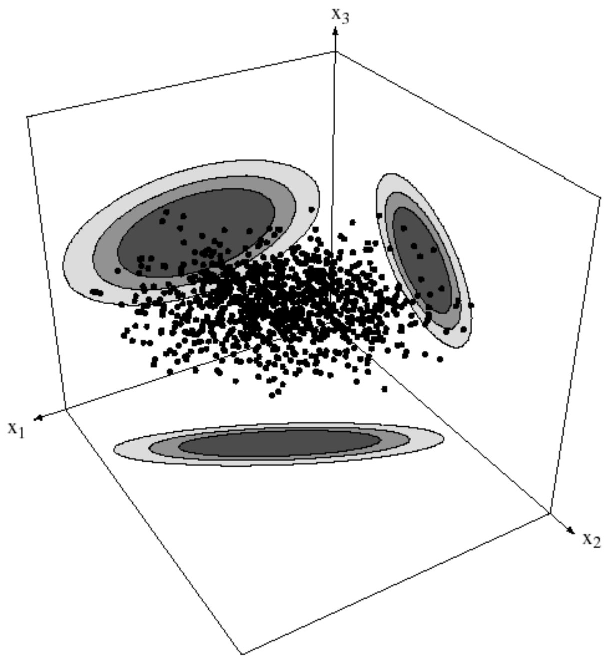
$p(A\backslash B) = \frac{p(A\cap B)}{p(B)}$ lub $p(A \cap B) = p(A\backslash B)p(B)$

- Prawdopodobieństwo apriori $p(A)$ jest specjalnym przypadkiem pradwopodobieństwa warunkowego $p(A\backslash)$ - gdy nie ma żadnej dodatkowej informacji.

- A prawdopodobieństwo B gdy A wyrazi się jako:

$$p(A \cap B) = p(B\backslash A)p(A)$$

Rozkład prawdopodobieństwa wielowymiarowy i niezależność



Dla x_1, x_3 $p(x_1, x_3) = p(x_1)p(x_3)$ - zmienne są niezależne. Pozostałe są zależne. Dwie zmienne x_i i x_j $i \neq j$ są statystycznie niezależne, jeśli

$$p(x_i, x_j) = p(x_i)p(x_j)$$

Reguła Bayesa

- $$p(A\backslash B) = \frac{p(B\backslash A)p(A)}{p(B)}$$
- Reguła Bayesa uwzględniająca dodatkową informację Y

$$p(A\backslash B, Y) = \frac{p(B\backslash A, Y)p(A\backslash Y)}{p(B\backslash Y)}$$

Interpretacja reguły Bayesa

- Dla modelu opisanego parametrami Q prawdopodobieństwo $p(Q\backslash x)$ określa wiarygodność modelu dla danych x

$$p(x\backslash Q) = \frac{p(Q\backslash x)p(x)}{p(Q)}$$

$$posteriori = \frac{wiarygodność \cdot apriori}{normalizacja}$$

- Dla *normalizacji =1* ("model evidence")

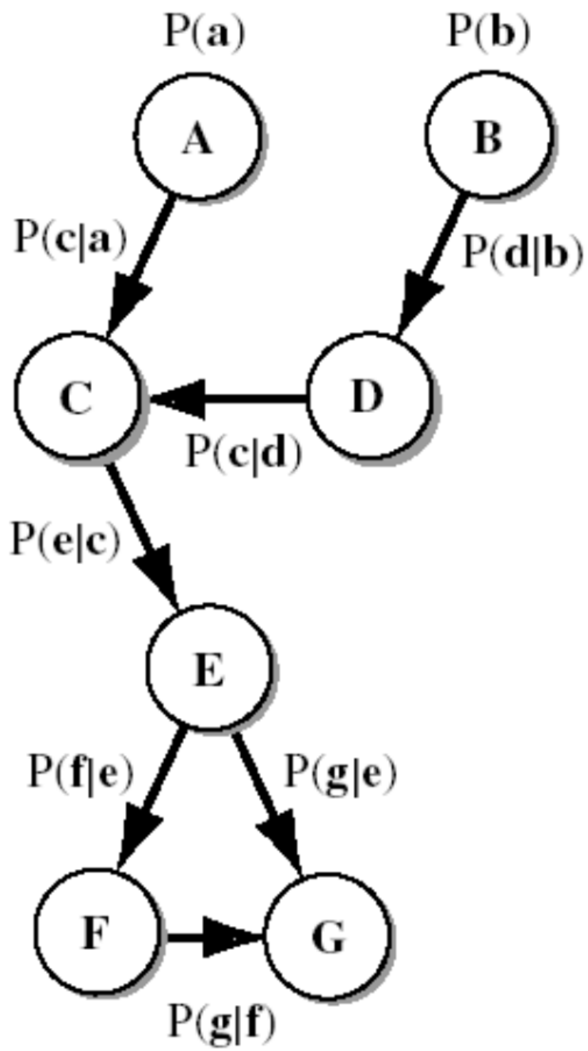
$$aposteriori \Leftarrow wiarygodność \cdot apriori$$

Sieci Bayesowskie

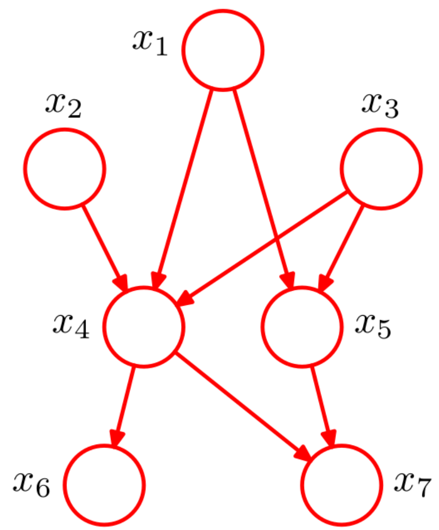
Sieć Bayesowska jest grafem, który:

- Zbiór zmiennych losowych $X, Y, \ldots$ jest reprezentowany poprzez węzły,
- Krawędzie (strzałki) łączą poszczególne węzły.
  Znaczenie strzałki pomiędzy węzłem X i Y:
  X ma bezpośredni wpływ na Y.
- Każdy węzeł posiada tablice prawdopodobieństw warunkowych swoich poprzedników.
- Graf nie posiada cykli określonych przez strzałki.

Sieć Bayesowska -przykład



Rozkład łączny a sieć Bayesa



- Łączny rozkład prawdopodobieństwa można zapisać jako iloczyn rozkładów warunkowych, po jednym dla każdego węzła:

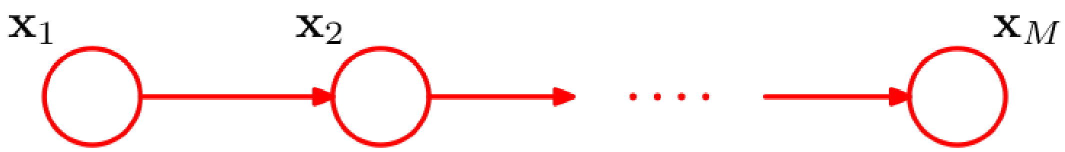
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

- Dla grafu z K węzłami rozkład łączny wyniesie:

$$p(x) = \Pi_{k=1}^{K} p(x_k|pa_k)$$

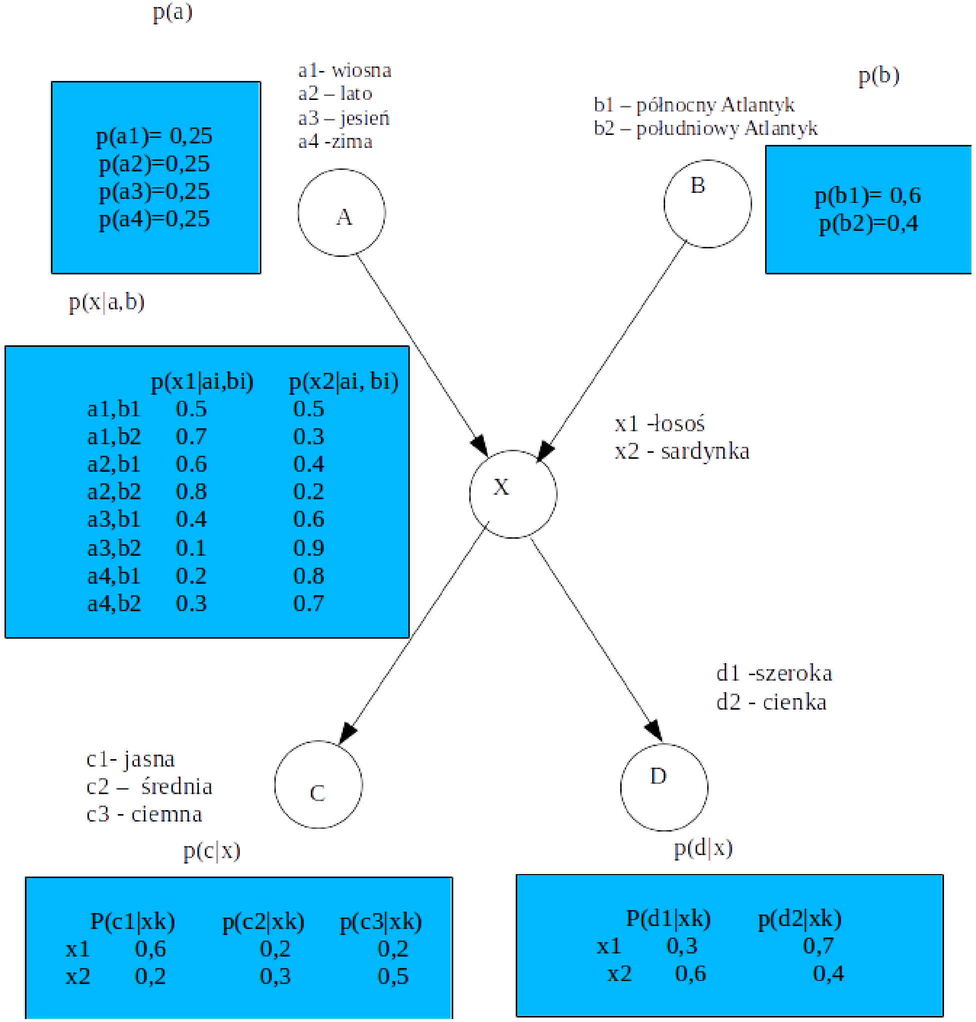
gdzie pa_k oznacza zbiór rodziców x_k.

Redukcja złożoności rozkładu łącznego



- Łańcuch M dyskretnych węzłów, każdy mający K stanów, wymaga określenia $M \cdot (K-1)$ parametrów, które rośną liniowa z M
- W pełni połączony wykres węzłów M (klika) miałby parametry $K^M - 1$, które rosną wykładniczo wraz z M.

Sieć Bayesowska - modelowanie zależności



Obliczenie rozkładów łącznych

- Jakie jest prawdopodobieństwo tego, że złowiona w *lecie* ryba jest z *północnego Atlantyku*, jest *sardynką* oraz jest *cienka* i *ciemna*?
- $p(a_3, b_1, x_2, c_3, d_2) = p(a_3)p(b_1)p(x_2|a_3, b_1)p(c_3|x_2)p(d_2|x_2) =$
- $$= 0,25 \cdot 0,6 \cdot 0,6 \cdot 0,5 \cdot 0,4 = 0,018$$

Jakie jest prawdopodobieństwo x_1

gdy wiadomo że, ryba jest jasna (c_1) i z północnego Atlantyku (b_2).

- Zakładając, że x_1, c_1, b_2 są niezależne to wówczas:

  $p(x_1|c_1, b_2) = \frac{p(x_1 \cap (c_1, b_2))}{p(c_1, b_2)} = \frac{p(x_1, c_1, b_2)}{p(c_1, b_2)}$

- zapisując:

  $p(x_1|c_1, b_2) = \alpha \sum_{a,d} p(x_1, \mathbf{a}, b_2, c_1, \mathbf{d})$

  gdzie α jest współczynnikiem normującym.

- otrzymujemy:

  $\alpha \sum_{a,d} p(x_1, \mathbf{a}, b_2, c_1, \mathbf{d}) =$

  $\alpha \sum_{\mathbf{a},\mathbf{d}} p(\mathbf{a}) p(b_2) p(x_1|\mathbf{a}, b_2) p(c_1|x_1) p(\mathbf{d}|x_1))$

- $\alpha \sum_{\mathbf{a},\mathbf{d}} p(\mathbf{a})p(b_2)p(x_1|\mathbf{a}, b_2)p(c_1|x_1)p(\mathbf{d}|x_1)) =$
  $= \alpha p(b_2)p(c_1|x_1) \cdot$
  $[\sum_{\mathbf{a}} p(\mathbf{a})p(x_1|\mathbf{a}, b_2)][\sum_{\mathbf{d}} p(\mathbf{d}|x_1)]$
- $= \alpha p(b_2)p(c_1|x_1)$
  $\cdot [p(a_1)p(x_1|a_1, b_2) + p(a_2)p(x_1|a_2, b_2) + p(a_3)p(x_1|a_3, b_2) +$
  $p(a_4)p(x_1|a_4, b_2)]$
  $\cdot [p(d_1|x_1) + p(d_2|x_1)]$
- $=$
  $\alpha(0.4)(0.6)[(0.25)(0.7) + (0.25)(0.8) + (0.25)(0.1) + (0.25)(0.3)]1.0$
- $= \alpha 0.114$

Uwaga do obliczeń

Zauważmy, że czynnik:

$$\sum_{\mathbf{d}} p(\mathbf{d}|x_1) = 1$$

nie wpływa na wynik obliczeń (porównaj kierunek strzałek)

Obliczenie α -normalizacja

- Obliczając w podobny sposób otrzymujemy $p(x_2|c_1, b_2) = \alpha 0.066$
- obliczone: $p(x_1|c_1, b_2) = \alpha 0.144$
- $\alpha p(x_1|c_1, b_2) + \alpha p(x_2|c_1, b_2) = 1$
- Skąd możemy obliczyć α: ($\alpha = 4,76$)
- $p(x_1|c_1, b_2) = 0,69$ i $p(x_2|c_1, b_2) = 0,31$

Jakie jest prawdopodobieństwo x_1 bez dodatkowej informacji ?

Należy obliczyć prawdopodobieństwo

- $p(x_1) = \sum_{i,j,k,l} p(x_1, a_i, b_j, c_k, d_l)$
  $= \sum_{i,j,k,l} p(a_i)p(b_j)p(x_1|a_i, b_j)p(c_k|x_1)p(d_l|x_1)$
- $= \sum_{i,j} p(a_i)p(b_j)p(x_1|a_i, b_j)$
- $= 0,25 \sum_{i,j} p(b_j)p(x_1|a_i, b_j)$
- $= 0,25[0,6(0,5 + 0,6 + 0,4 + 0,2) + 0,4(0,7 + 0,8 + 0,1 + 0,3)] = 0,445$

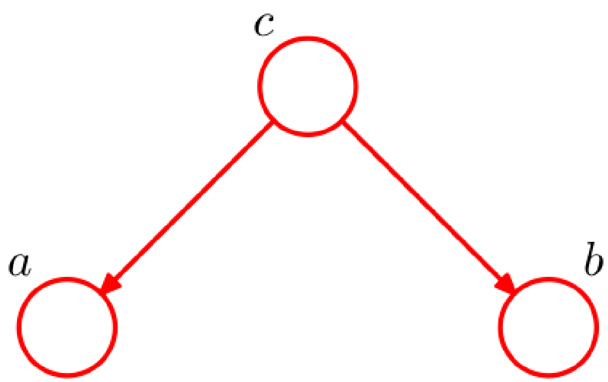
Reguła naiwnego Bayesa (ang. *naive Bayes*)

Gdy zależność pomiędzy zmiennymi a, b jest nieznana (zakłada się, że zmienne są niezależne) wówczas prawdopodobieństwo warunkowe przejmie postać:

$$p(x|a, b) = p(x|a)p(x|b)$$

- Naive Bayes jest prostszy w użyciu,
- Bardzo często sprawdza się.

Warunkowa niezależność



- Załóżmy, że rozkład warunkowy a pod warunkiem b i c nie zależy od wartości b,

$$p(a|b,c) = p(a|c)$$

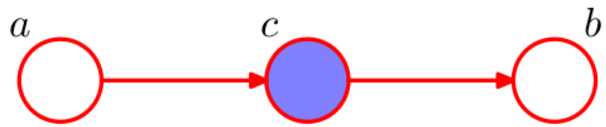
- Rozkład łączny a i b uwarunkowany przez c wynosi

$$p(a,b|c) = p(a|b,c)p(b|c) = p(a|c)p(b|c)$$

- Fakt, że a jest niezależne od b dla danego c będziemy zapisywać jako:

$$a \perp b|c$$

Warunkowa niezależność - przypadek 1



- Gdy wiemy, że zaszło c wówczas:

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(b|c)p(c|a)p(a)}{p(c)} = p(a|c) \cdot p(b|c)$$

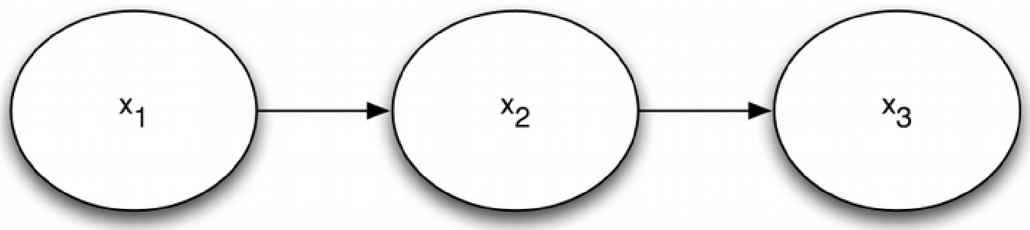
co oznacza, że dla zmienne a i b są niezależne: $a \perp b|c$

- Gdy nie mamy żadnej informacji o c:

$$p(a,b) = p(a) \sum_c p(c|a)p(b|c) = p(a) \cdot p(b|a)$$

co oznacza, że dla zmiennych a i b nie zachodzi niezależność: $a \not\perp b|c$

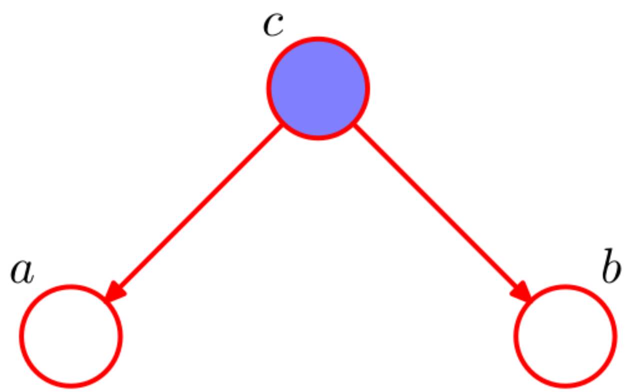
Przypadek 1: Łańcuch Markowa



- Przykładem takiego łańcucha jest proces lokalizacji robota,
- Jeśli robot zna aktualny stan, x_2, to nie potrzebuje informacji o przeszłych stanach x_1, w celu określenia następnego stanu x_3.
- Można stwierdzić, że x_1 i x_3 są niezależne, jeśli x_2 jest znane.

$$P(x_3|x_2, x_1) = P(x_3|x_2)$$

Warunkowa niezależność - przypadek 2



- Gdy wiemy, że zaszło c wówczas

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(a|c)p(b|c)p(c)}{p(c)} = p(a|c) \cdot p(b|c)$$

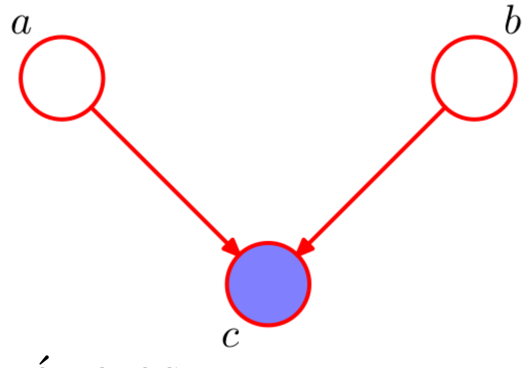
  co oznacza, że dla zmienne a i b są niezależne: $a \perp b|c$

- Gdy nie mamy żadnej informacji o c:

$$p(a,b) = \sum_c p(a|c)p(b|c)p(c)$$

  co oznacza, że dla zmiennych a i b nie zachodzi niezależność: $a \not\perp b|c$

Warunkowa niezależność - przypadek 3



- Gdy wiemy, że zaszło c wówczas

$$p(a,b|c) = \frac{p(a,b,c)}{p(c)} = \frac{p(c|a,b)p(b)p(a)}{p(c)}$$

co oznacza, że dla zmiennych a i b nie zachodzi niezależność: $a \not\perp b|c$

- Gdy nie mamy żadnej informacji o c:

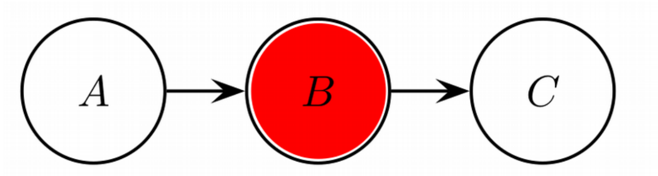
$$p(a,b) = \sum_c p(c|a,b)p(a)p(b) = p(a) \cdot p(b)$$

co oznacza, że dla zmienne a i b są niezależne: $a \perp b|c$

Koncepcje d-separacji

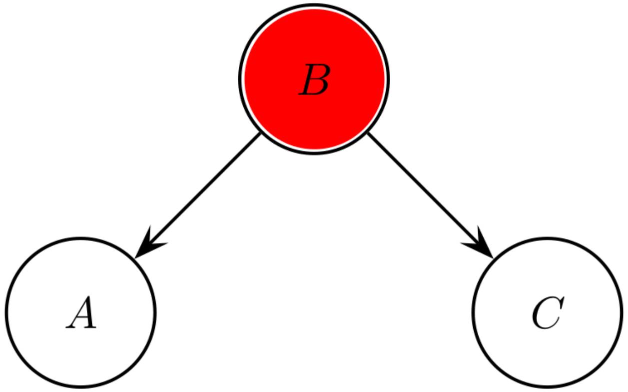
- W przypadku grafów kierunkowanych możliwe jest sprawdzenie, czy dana właściwość niezależności warunkowej utrzymuje się, stosując test graficzny zwany d-separacją.
- Wymaga to sprawdzenia, czy ścieżki łączące dwa zestawy węzłów zostały „zablokowane”.

Blokada węzła



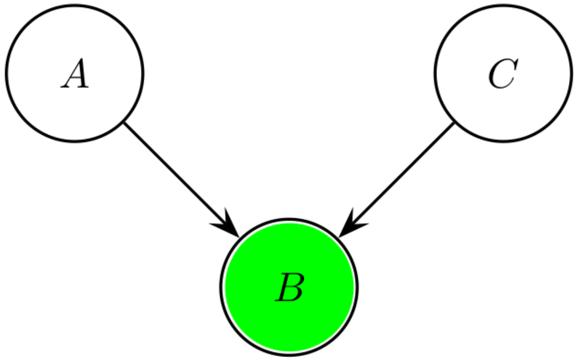
- W łańcuchu węzłów, C jest niezależny od A, jeśli B jest znany. Oznacza to, że na każdej ścieżce przechodzącej przez łańcuch Markowa występuje blokada za pomocą znanego węzła środkowego.
- Gdy x_2 nie jest znane, znajomość stanów przeszłych może dostarczyć informacji na temat aktualnego stanu x_3.
- Tak samo jest z powiedzeniem że x_1 i x_3 mogą być zależne, jeśli x_2 nie jest znane

Przypadek 2: Jeden rodzić, dwoje dzieci



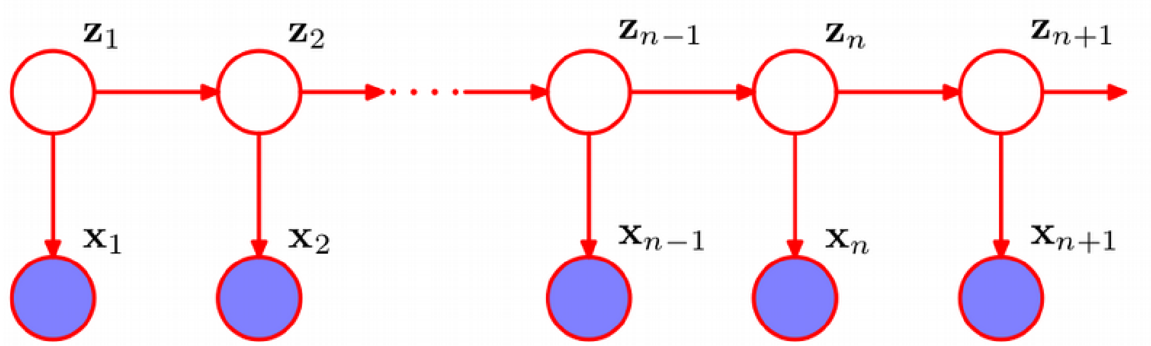
- A jest niezależny od C przy znanym B.
- Oznacza to, że istnieje blokada na ścieżce od A do C jeśli rodzic B jest znany.

Przypadek 3: Dwoje rodziców, jedno dziecko



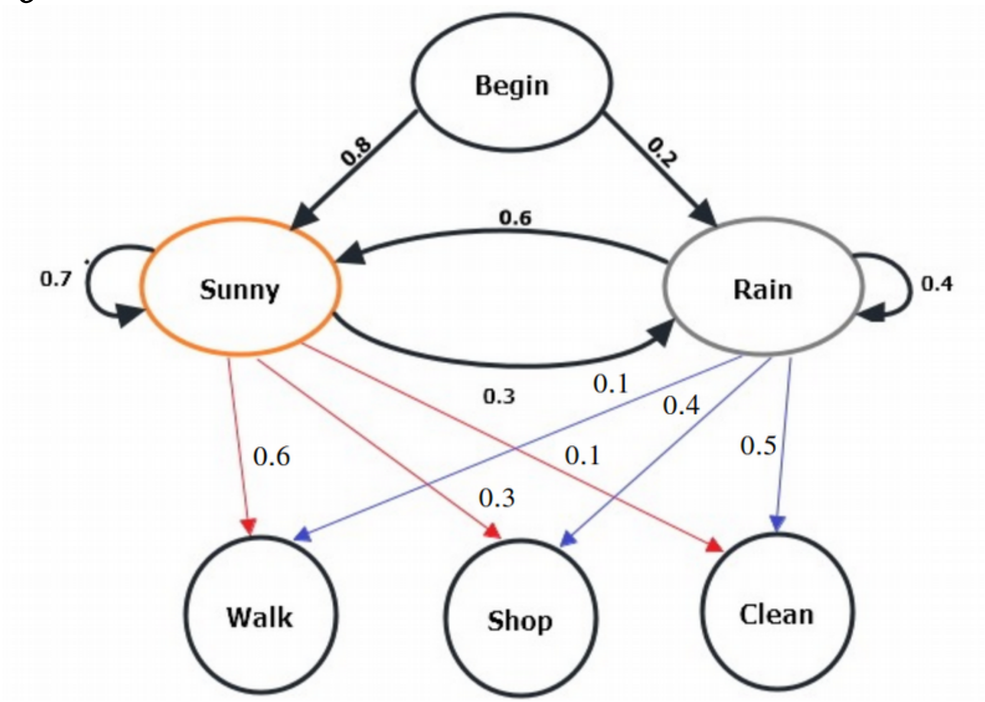
- A jest zależne od C, jeśli B jest znane. Odwrotność jest również prawdziwa, A jest niezależne od C, jeśli B nie jest znane.
- Istnieje blokada na ścieżce przechodzącej przez przypadek "dwóch rodziców, jedno dziecko", jeśli B jest nieznany.

Ukryte Modele Markowa (ang. Hidden Markov Model)



- Ukryty model Markowa można postrzegać jako specyficzny przykład modelu przestrzeni stanów, w którym zmienne ukryte są dyskretne.
- HMM jest szeroko stosowany w rozpoznawaniu mowy, modelowaniu języka naturalnego, rozpoznawaniu pisma ręcznego online.

HMM przykład



- Możliwe są 3 zajęcia: spacer, zakupy i sprzątanie.
- Wybór, zależy wyłącznie od pogody danego dnia - *macierz prawdopodobieństwa emisji.*
- Pogoda na następny dzień zależy od pogody z poprzedniego dnia - *macierz prawdopodobieństwa przejścia.*

Kod HMM w pythonie

```
start = ['R','Su']
#prawdopodobienstwo startu
p_start = [0.2,0.8]
#macierz przejscia
t1 = ['R','Su']
p_t1=[[0.4,0.6],[0.3,0.7]]
t2 = ['Walk','Shop','Clean']
#macierz emisji
p_t2=[[0.1,0.4,0.5],[0.6,0.3,0.1]]
initial = np.random.choice(start,replace=True,p=p_start)
n = 20
st = 1
for i in range(n):
    if st:
        state = initial
        st = 0
        print(state)
    if state == 'R':
        activity = np.random.choice(t2,p=p_t2[0])
        print(state)
        print(activity)
        state = np.random.choice(t1,p=p_t1[0])
    elif state == 'Su':
        activity = np.random.choice(t2,p=p_t2[1])
        print(state)
        print(activity)
        state = np.random.choice(t1,p=p_t1[1])
    print("\n")
```

Projekt 1

1. Napisz program, uczący się gry “papier, kamień, nożyce”. Jako algorytm uczący zastosuj łańcuch Markowa z macierzą przejść pomiędzy trzema stanami (Papier, Kamień, Nożyce). Nauka gry polega na korekcie macierzy przejść (prawdopodobieństwa warunkowe zmiany stanu),
2. Wartość wypłaty: 1 w przypadku wygranej, -1 w przypadku przegranej i 0 w przypadku remisu,
3. Przeprowadź ciąg kilkudziesięciu gier “kamień, papier, nożyce”. Sporząź wykres jak zmienia się stan kasy w każdym kroku gry.