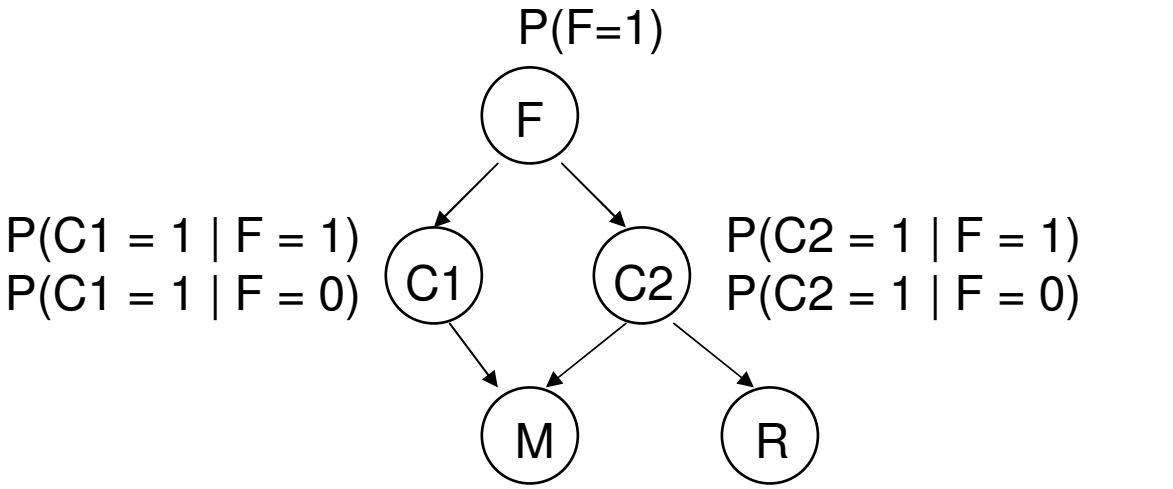
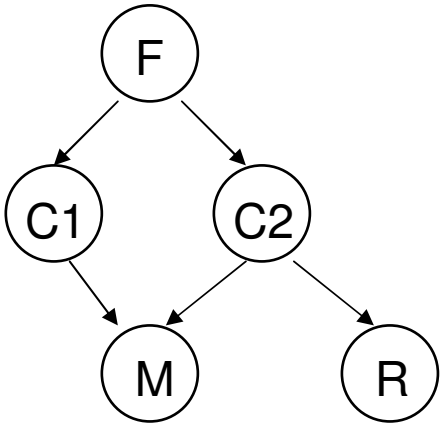
Sieci bayesowskie



F	pali papierosy
C1	chory na chorobę 1
C2	chory na chorobę 2
M	łatwo się męczy
R	wynik RTG płuc

Sieci bayesowskie



Prawdopodobieństwo łączne:

$$P(C1, C2, M, R, F) = P(F) * P(C1 \mid F) * P(C2 \mid F) * P(M \mid C1, C2) * P(R \mid C2)$$

$$P(C2=1 \mid R=1, F=0) = \frac{P(C2=1, R=1, F=0)}{P(R=1, F=0)} = \frac{\sum_{i=0}^{1}\sum_{j=0}^{1} P(C1=i, C2=1, M=j, R=1, F=0)}{\sum_{i=0}^{1}\sum_{j=0}^{1}\sum_{k=0}^{1} P(C1=i, C2=j, M=k, R=1, F=0)}$$

Sieci bayesowskie - zalety

- Czytelna reprezentacja wiedzy o zależnościach przyczynowych
- Oszczędna reprezentacja łącznego rozkładu prawdopodobieństwa
- Efektywne algorytmy wnioskowania

Sieci bayesowskie

1. Uczenie sieci
   - uczenie struktury sieci
   - szacowanie parametrów
2. Wnioskowanie na podstawie sieci

Uczenie parametrów

Założenie:

Znana jest struktura sieci bayesowskiej

Zadanie:

Wyznaczenie prawdopodobieństw warunkowych

Uczenie parametrów

1. Ocena częstości na podstawie pierwotnej wiedzy (opartej na naszych subiektywnych oszacowaniach lub pochodzącej z innych źródeł).

2. Uaktualnienie prawdopodobieństw na podstawie danych.

Uczenie parametrów, zmienna binarna – rozkład beta

Do przedstawienia naszej pierwotnej wiedzy o względnej częstości nadaje się funkcja gęstości beta:

$$\rho(f) = \frac{\Gamma(N)}{\Gamma(a)\Gamma(b)} f^{a-1}(1-f)^{b-1},$$

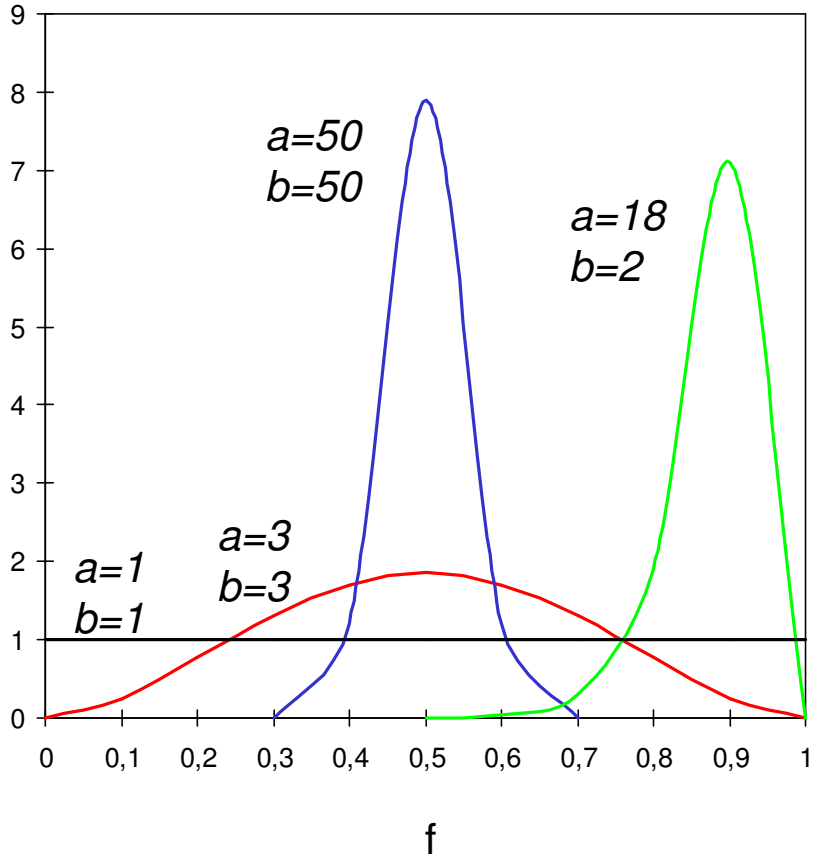
$$N = a + b$$

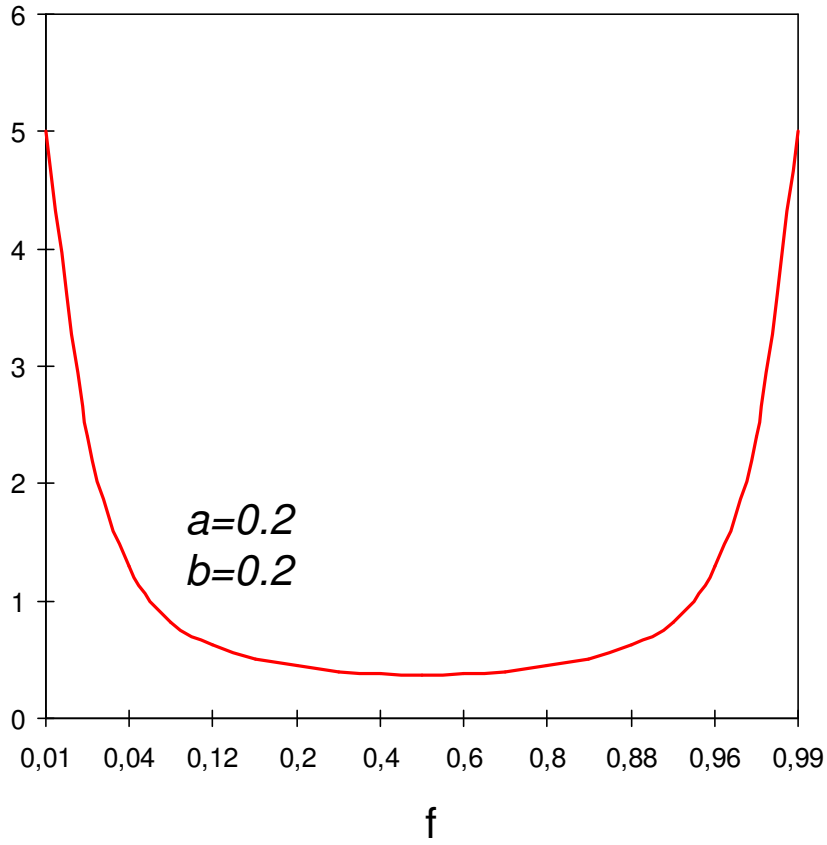
f – względna częstość
a, *b*, *N* – parametry rozkładu

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

$$x \geq 1 \text{ i } x \in C \Rightarrow \Gamma(x) = (x - 1)!$$

Uczenie parametrów, zmienna binarna - rozkład beta





Uczenie parametrów, zmienna binarna

Oznaczenia:

- *X* - zmienna losowa binarna przyjmująca wartości 1 i 2

- *F* - zmienna losowa reprezentująca naszą wiedzę dotyczącą względnej częstości z jaką zmienna *X* przyjmuje wartość 1

- *f* – konkretna wartość *F*

Uczenie parametrów, zmienna binarna

- Jeżeli F ma rozkład beta z parametrami a, b, $N = a + b$, to wartość oczekiwana $E(F) = a / N$;

- Jeżeli zmienna X przyjmuje wartości 1 lub 2 i $P(X = 1 \mid f) = f$, to $P(X = 1) = E(F)$;

- Jeśli F ma rozkład beta, to $P(X = 1) = a / N$;

Przykład: rzut monetą

przypuszczamy, że na 100 rzutów monetą ok. 50 razy wypadnie orzeł, zatem zakładamy rozkład beta(f; 50,50), stąd

$P(X = \text{orzeł}) = 50 / (50 + 50) = 0.5$

Uczenie parametrów, zmienna binarna – uaktualniania na podstawie danych

Dane:

$\rho(f)$ – rozkład gęstości zmiennej losowej F reprezentującej względną częstość

d - zbiór danych zawierający M przykładów

s – liczba wystąpień pierwszej wartości w zbiorze d

t - liczba wystąpień drugiej wartości w zbiorze d ($M = s + t$)

Problem:

Jak uaktualnić na podstawie danych d naszą pierwotną wiedzę reprezentowaną przez $\rho(f)$?

Rozwiązanie:

Jeżeli $\rho(f)$ = beta(f; a, b), to $\boldsymbol{\rho(f \mid d) = \text{beta}(f; a + s, b + t)}$.

Uczenie parametrów, zmienna binarna – uaktualniania na podstawie danych

Przykład:

Rzut oszczepem wykonany 10 razy (może wbić się w ziemię lub wylądować płasko).

Pierwotna wiedza: beta(f; 3, 3).

Wyniki rzutów (1-wbity): d={1,1,2,1,1,1,1,1,2,1}

$a = b = 3$, $s = 8$, $t = 2$

$\rho(f \mid d) = \text{beta}(f; 3 + 8, 3 + 2) = \text{beta}(f; 11, 5)$

$P(X = 1 \mid d) = E(F \mid d) = (a + s) / (N + M) = (3 + 8) / (6 + 10) = 0.69$

Wybór wartości *a* oraz *b*

- ***a* = *b* = 1**
  brak jakiejkolwiek wiedzy na temat względnej częstości lub dla zachowania obiektywizmu; zakładamy, że wszystkie wartości z przedziału [0, 1] są jednakowo prawdopodobne;

- ***a*, *b* > 1**
  prawdopodobnie względna częstość wystąpienia pierwszej z dopuszczalnych wartości zmiennej wynosi ok. $a/(a+b)$; im większe przekonanie, tym większe wartości *a*, *b*;

- ***a*, *b* < 1**
  prawdopodobnie względna częstość wystąpienia jednej z wartości zmiennej jest bardzo mała;

Uczenie parametrów - *przykład 1*

- trzy urny zawierające kule z numerami 1 i 2;
- losujemy kulę z pierwszej urny. Jeśli wylosowawo 1, to losujemy z urny drugiej, w przeciwnym razie z trzeciej;
- X_1, X_2 – zmienne reprezentujące wyniki dwóch losowań;
- beta(f_{11}; 1, 1) reprezentuje pierwotną wiedzę na temat częstości wylosowania kuli 1 przy pierwszym losowaniu;
- beta(f_{21}; 1, 1) reprezentuje pierwotną wiedzę na temat częstości wylosowania kuli 1 przy drugim losowaniu, jeżeli przy pierwszym wylosowano 1;
- beta(f_{22}; 1, 1) reprezentuje pierwotną wiedzę na temat częstości wylosowania kuli 1 przy drugim losowaniu, jeżeli przy pierwszym wylosowano 2;

$P(X_1 = 1) = ½$

$P(X_2 = 1 \mid X_1 = 1) = ½$
$P(X_2 = 1 \mid X_1 = 2) = ½$

np.:
$P(X_1 = 2, X_2 = 1) = P(X_2 = 1 \mid X_1 = 2)\ P(X_1 = 2) = ½ * ½ = 1/4$

Uczenie parametrów - *przykład 1*

Dane *d*:

X_1	X_2
1	1
1	1
1	1
1	2
2	1
2	1
2	2

- $\rho(f_{11}|d) = beta(f_{11}; 1 + 4, 1 + 3)$
- $\rho(f_{21}|d) = beta(f_{21}; 1 + 3, 1 + 1)$
- $\rho(f_{22}|d) = beta(f_{22}; 1 + 2, 1 + 1)$

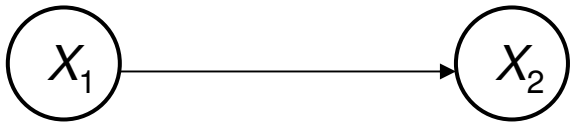
$P(X_1 = 1) = 5/9$

$P(X_2 = 1 \mid X_1 = 1) = 2/3$
$P(X_2 = 1 \mid X_1 = 2) = 3/5$

np.:
$P(X_1 = 2, X_2 = 1) = P(X_2 = 1 \mid X_1 = 2)\ P(X_1 = 2) = 3/5 * 4/9 = 4/15$

Uczenie parametrów - *przykład 2*

- X_1 – zmienna mówiąca czy pacjent pali papierosy (1-tak, 2-nie)
- X_2 – zmienna mówiąca czy pacjent ma chore płuca (1-tak, 2-nie)
- $\rho(f_{11}) = beta(f_{11}; 1, 1)$
- $\rho(f_{21}) = beta(f_{21}; 1, 1)$
- $\rho(f_{22}) = beta(f_{22}; 1, 1)$



$P(X_1 = 1) = ½$

$P(X_2 = 1 \mid X_1 = 1) = ½$
$P(X_2 = 1 \mid X_1 = 2) = ½$

np.:

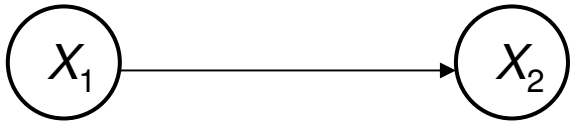
$P(X_2 = 1) = P(X_2 = 1, X_1 = 1) + P(X_2 = 1, X_1 = 2) =$
$= P(X_2 = 1 \mid X_1 = 1)\, P(X_1 = 1) + P(X_2 = 1 \mid X_1 = 2)\, P(X_1 = 2) =$
$= ½ * ½ + ½ * ½ = ½$

Uczenie parametrów - *przykład 2*

Dane *d*:

X_1	X_2
1	2
1	1
2	1
2	2
2	1
2	1
1	2
2	2

- $\rho(f_{11}|d) = beta(f_{11}; 1 + 3, 1 + 5)$
- $\rho(f_{21}|d) = beta(f_{21}; 1 + 1, 1 + 2)$
- $\rho(f_{22}|d) = beta(f_{22}; 1 + 3, 1 + 2)$



$P(X_1 = 1) = 2/5$

$P(X_2 = 1 \mid X_1 = 1) = 2/5$
$P(X_2 = 1 \mid X_1 = 2) = 4/7$

np.:

$P(X_2 = 1) = P(X_2 = 1, X_1 = 1) + P(X_2 = 1, X_1 = 2) =$
$= P(X_2 = 1 \mid X_1 = 1)\, P(X_1 = 1) + P(X_2 = 1 \mid X_1 = 2)\, P(X_1 = 2) =$
$= 2/5 * 2/5 + 4/7 * 3/5 = 0.50286$

Uczenie parametrów - *przykład 2*

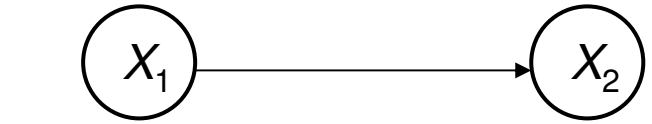
- Początkowo $P(X_2 = 1) = 0.5$
- Wśród danych uczących były 4 przykłady, dla których $X_2 = 1$ oraz 4 przykłady, dla których $X_2 = 2$
- Po uaktualnieniu parametrów $P(X_2 = 1) = 0.50286$

?

Uczenie parametrów, odpowiedni rozmiar danych - *przykład 2*

Początkowa wiedza:

- $\rho(f_{11}) = beta(f_{11}; \mathbf{2}, \mathbf{2})$
- $\rho(f_{21}) = beta(f_{21}; 1, 1)$
- $\rho(f_{22}) = beta(f_{22}; 1, 1)$



$P(X_1 = 1) = ½$ $P(X_2 = 1 \mid X_1 = 1) = ½$
$P(X_2 = 1 \mid X_1 = 2) = ½$

Po uaktualnieniu parametrów:

- $\rho(f_{11}|d) = beta(f_{11}; \mathbf{2} + 3, \mathbf{2} + 5)$
- $\rho(f_{21}|d) = beta(f_{21}; 1 + 1, 1 + 2)$
- $\rho(f_{22}|d) = beta(f_{22}; 1 + 3, 1 + 2)$

$P(X_1 = 1) = \mathbf{5/12}$ $P(X_2 = 1 \mid X_1 = 1) = 2/5$
$P(X_2 = 1 \mid X_1 = 2) = 4/7$

np.:

$$P(X_2 = 1) = P(X_2 = 1, X_1 = 1) + P(X_2 = 1, X_1 = 2) =$$
$$= P(X_2 = 1 \mid X_1 = 1)\, P(X_1 = 1) + P(X_2 = 1 \mid X_1 = 2)\, P(X_1 = 2) =$$
$$= 2/5 * 5/12 + 4/7 * 7/12 = 0.5$$

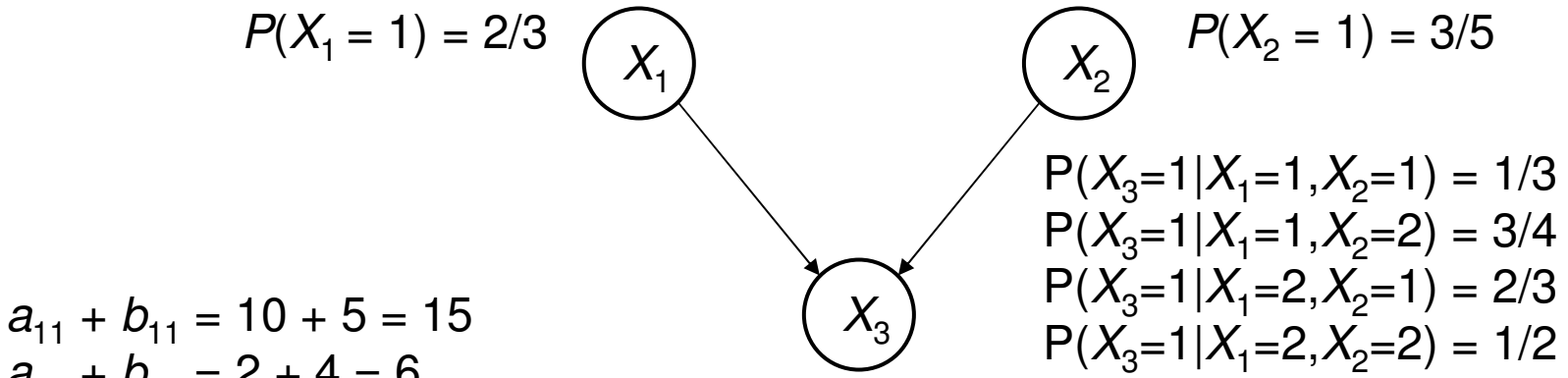
Uczenie parametrów, odpowiedni rozmiar danych

Definicja:
Dla każdego *i* oraz *j* funkcje gęstości są postaci $beta(f_{ij}; a_{ij}, b_{ij})$.
Jeżeli istnieje liczba *N* taka, że
$N_{ij} = a_{ij} + b_{ij} = P(pa_{ij}) * N$
to mówimy, że sieć ma odpowiedni rozmiar danych *N*.
$P(pa_{ij})$ – prawd. tego, że węzły rodzice danego węzła były w stanie *ij*

Dla danych z *przykładu 2*:
$N_{21} = 2 + 3 = 5/12 * 12$
$N_{22} = 4 + 3 = 7/12 * 12$
$N = 12$

Uczenie parametrów, odpowiedni rozmiar danych

- $beta(f_{11}; 10, 5)$ $\quad f_{11} = P(X_1=1|f_{11})$
- $beta(f_{21}; 9, 6)$ $\quad f_{21} = P(X_2=1|f_{21})$
- $beta(f_{31}; 2, 4)$ $\quad f_{31} = P(X_3=1|X_1=1,X_2=1,f_{31})$
- $beta(f_{32}; 3, 1)$ $\quad f_{32} = P(X_3=1|X_1=1,X_2=2,f_{32})$
- $beta(f_{33}; 2, 1)$ $\quad f_{33} = P(X_3=1|X_1=2,X_2=1,f_{33})$
- $beta(f_{34}; 1, 1)$ $\quad f_{34} = P(X_3=1|X_1=2,X_2=2,f_{34})$



$a_{11} + b_{11} = 10 + 5 = 15$

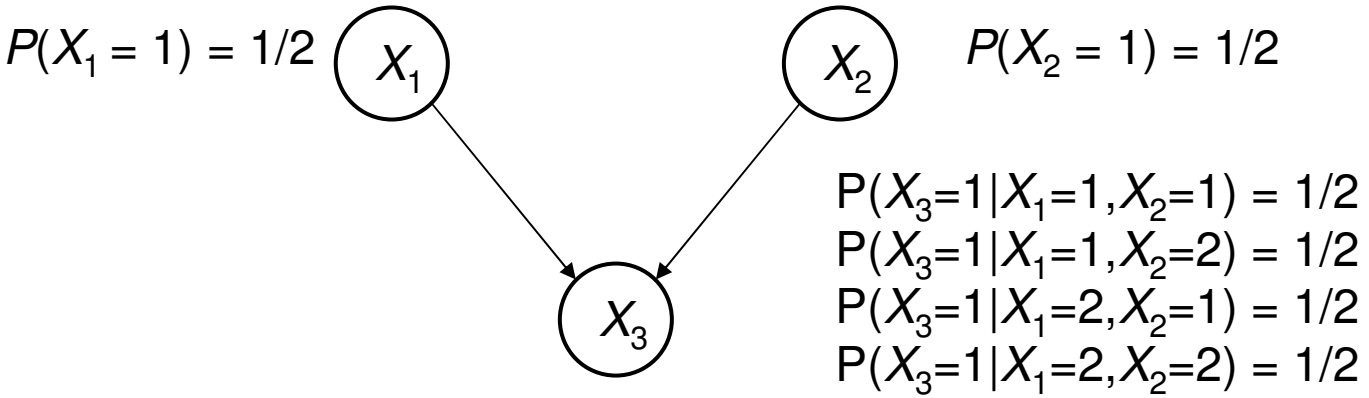
$a_{31} + b_{31} = 2 + 4 = 6$

$P(\text{pa}_{31}) * N = P(X_1 = 1, X_2 = 1) * N = 2/3 * 3/5 * 15 = 6$

Uczenie parametrów, odpowiedni rozmiar danych

W jaki sposób wyrazić brak początkowej wiedzy (a=b=1) zachowując odpowiedni rozmiar danych?

- $beta(f_{11}; 1, 1)$ $\quad f_{11} = \mathrm{P}(X_1=1|f_{11})$
- $beta(f_{22}; 1, 1)$ $\quad f_{22} = \mathrm{P}(X_2=1|f_{22})$
- $beta(f_{31}; ¼, ¼)$ $\quad f_{31} = \mathrm{P}(X_3=1|X_1=1,X_2=1,f_{31})$
- $beta(f_{32}; ¼, ¼)$ $\quad f_{32} = \mathrm{P}(X_3=1|X_1=1,X_2=2,f_{32})$
- $beta(f_{33}; ¼, ¼)$ $\quad f_{33} = \mathrm{P}(X_3=1|X_1=2,X_2=1,f_{33})$
- $beta(f_{34}; ¼, ¼)$ $\quad f_{34} = \mathrm{P}(X_3=1|X_1=2,X_2=2,f_{34})$



Uczenie parametrów, brakujące wartości

Początkowa wiedza:

- $\rho(f_{11}) = beta(f_{11}; 2, 2)$
- $\rho(f_{21}) = beta(f_{21}; 1, 1)$
- $\rho(f_{22}) = beta(f_{22}; 1, 1)$

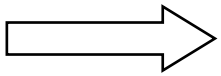
$P(X_1 = 1) = 1/2$

$P(X_2 = 1 \mid X_1 = 1) = 1/2$
$P(X_2 = 1 \mid X_1 = 2) = 1/2$

Dane *d*:

X_1	X_2
1	1
1	**?**
1	1
1	2
2	**?**

$P(X_2 = 1 \mid X_1 = 1) = ½$
$P(X_2 = 1 \mid X_1 = 2) = ½$



Dane *d'*:

X_1	X_2	*Liczba wystąpień*
1	1	1
1	**1**	½
1	**2**	½
1	1	1
1	2	1
2	**1**	½
2	**2**	½

Uczenie parametrów, brakujące wartości

Początkowa wiedza:

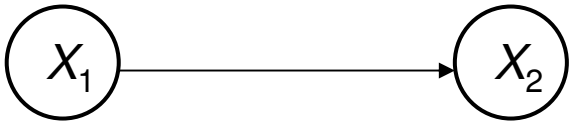
- $\rho(f_{11}) = beta(f_{11}; 2, 2)$
- $\rho(f_{21}) = beta(f_{21}; 1, 1)$
- $\rho(f_{22}) = beta(f_{22}; 1, 1)$

Po uaktualnienie parametrów

- $\rho(f_{11}|d') = beta(f_{11}; 2 + 4, 2 + 1)$
- $\rho(f_{21}|d') = beta(f_{21}; 1 + 5/2, 1 + 3/2)$
- $\rho(f_{22}|d') = beta(f_{22}; 1 + 1/2, 1 + 1/2)$

Dane d':

X_1	X_2	*Liczba wystąpień*
1	1	1
1	**1**	½
1	**2**	½
1	1	1
1	2	1
2	**1**	½
2	**2**	½



$P(X_1 = 1) = 6 / (6 + 3) = 2/3$

$P(X_2 = 1 \mid X_1 = 1) = 7/2 / (7/2 + 5/2) = 7/12$
$P(X_2 = 1 \mid X_1 = 2) = 1/2$

Uczenie parametrów, brakujące wartości

Wyznaczenie najbardziej prawdopodobnego rozkładu (algorytm EM)

Powtarzaj na przemian k razy:

1. oblicz oczekiwane wartości s_{ij}' oraz t_{ij}' na podstawie rozkładu prawdopodobieństw i danych d (*expectation*)

2. oblicz wartości f_{ij}' na podstawie s_{ij}' oraz t_{ij}' (*maximization*)

Można w ten sposób wyznaczyć f maksymalizujące $\rho(f|d)$.
Problem: ryzyko znalezienia maksimum lokalnego.

Uczenie parametrów, zmienne dyskretne o dowolnej liczbie wartości – rozkład Dirichleta

$$\rho(f_1, f_2, \ldots, f_{r-1}) = \frac{\Gamma(N)}{\prod_{k=1}^{r} \Gamma(a_k)} f_1^{a_1-1} f_2^{a_2-1} \ldots f_r^{a_r-1}$$

$$0 \le f_k \le 1, \quad \sum_{k=1}^{r} f_k = 1, \quad \sum_{k=1}^{r} a_k = N$$

r – liczba różnych wartości (dla zmiennej binarnej $r = 2$ i mamy rozkład beta)
a_i – liczba wystąpień i-tej wartości

Uczenie parametrów, zmienne dyskretne – rozkład Dirichleta

Jeżeli zmienne F_1, F_2,...,F_r mają rozkład Dirichleta z parametrami a_1, a_2...,a_r, $a_1 + a_2 + ... + a_r = N$, to dla $1 \leq k \leq r$

$$E(F_k) = \frac{a_k}{N}$$

Jeżeli zmienna losowa X przyjmuje wartości 1,2,...,r; i dla zmiennych losowych F_1, F_2,..., F_r, $P(X = k|f_k) = f_k$, to

$$P(X = k) = \frac{a_k}{N}$$

Uczenie parametrów, zmienne dyskretne – rozkład Dirichleta

Przykład 1

Rzut kostką
$Dir(f_1, f_2, f_3, f_4, f_5; 50, 50, 50, 50, 50, 50)$
$P(X = 4) = 50 / 300 = 1/ 6$

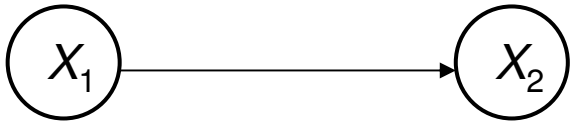
Przykład 2

Kolor skarpetek (białe, czarne, kolorowe)
$Dir(f_1, f_2; 2, 2, 4)$
$P(X = \text{białe}) = 2 / (2 + 2 + 4) = 1 / 4$
$P(X = \text{kolorowe}) = 4 / (2 + 2 + 4) = 1 / 2$

Uczenie parametrów

Zmienna X_1 przyjmuje 3 wartości, zmienna X_2 przyjmuje 4 wartości

- $\rho(f_{111}, f_{112}) = Dir(f_{111}, f_{112}; 4, 8, 10)$
- $\rho(f_{211}, f_{212}, f_{213}) = Dir(f_{211}, f_{212}, f_{213}; 1, 1, 1, 1)$
- $\rho(f_{221}, f_{222}, f_{223}) = Dir(f_{221}, f_{222}, f_{223}; 2, 4, 1, 1)$
- $\rho(f_{231}, f_{232}, f_{233}) = Dir(f_{231}, f_{232}, f_{233}; 1, 3, 4, 2)$



$P(X_1 = 1) = 2/11$
$P(X_1 = 2) = 4/11$
$P(X_1 = 3) = 5/11$

$P(X_2 = 1 \mid X_1 = 1) = ¼$
$P(X_2 = 2 \mid X_1 = 1) = ¼$
$P(X_2 = 3 \mid X_1 = 1) = ¼$
$P(X_2 = 4 \mid X_1 = 1) = ¼$

$P(X_2 = 1 \mid X_1 = 2) = ¼$
$P(X_2 = 2 \mid X_1 = 2) = ½$
$P(X_2 = 3 \mid X_1 = 2) = 1/8$
$P(X_2 = 4 \mid X_1 = 2) = 1/8$

$P(X_2 = 1 \mid X_1 = 3) = 1/10$
$P(X_2 = 2 \mid X_1 = 3) = 3/10$
$P(X_2 = 3 \mid X_1 = 3) = 2/5$
$P(X_2 = 4 \mid X_1 = 3) = 1/5$

Uczenie parametrów, zmienne dyskretne – uaktualnianie na podstawie danych

Dane:

$\rho(f_1,f_2,...,f_{r-1})$ – rozkład gęstości zmiennych reprezentujących względną częstość

d - zbiór danych zawierający M przykładów

s_i – liczba wystąpień i-tej wartości w zbiorze d

Problem:

Jak uaktualnić na podstawie danych d naszą pierwotną wiedzę reprezentowaną przez $\rho(f_1,f_2,...,f_{r-1})$?

Rozwiązanie:

Jeżeli $\rho(f_1, f_2,..., f_{r-1}) = Dir(f_1, f_2,..., f_{r-1}; a_1, a_2,..., a_r)$, to
$\boldsymbol{\rho(f_1, f_2,..., f_{r-1} \mid d) = Dir(f_1, f_2,..., f_{r-1}; a_1 + s_1, a_2 + s_2,..., a_r + s_r)}$.

Metody uczenia struktury sieci - podział

Ze względu na sposób optymalizacji:

- Dokładne (przeglądanie wszystkich możliwych sieci w celu wybrania optymalnej, wykonalne jedynie w przypadku małej liczby zmiennych).
- **Przybliżone** (nie gwarantują optymalnego rozwiązania, ale przeważnie prowadzą do rozwiązań bliskich optymalnemu).

Ze względu na wynik końcowy:

- **Wybór modelu** (wyznaczenie jednej sieci).
- Uśrednianie modelu (wyznaczenie wielu sieci a następnie uśrednianie prawdopodobieństw podczas wnioskowania).

Uczenie struktury sieci bayesowskiej – miara jakości

Im większe prawdopodobieństwo $P(G|d)$, tym lepsza sieć G (d – dane uczące).

$$P(G|d) = \frac{P(d|G)P(G)}{P(d)}$$

Miara jakości sieci G zbudowanej na podstawie danych d:

$$score_B(d,G) = P(d|G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i^{(G)}} \frac{\Gamma(N_{ij}^{(G)})}{\Gamma(N_{ij}^{(G)} + M_{ij}^{(G)})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk}^{(G)} + s_{ijk}^{(G)})}{\Gamma(a_{ijk}^{(G)})}$$

n - liczba węzłów (zmiennych)

q_i – liczba kombinacji wartości przyjmowanych przez rodziców węzła i-tego

r_i – liczba wartości przyjmowanych przez zmieną (węzeł) i-tą

M_{ij} – liczba przykładów uczących, dla których rodzice węzła i-tego przyjmują j-tą kombinację wartości

s_{ijk} – liczba przykładów uczących, dla których zmienna i-ta przyjmuje wartość k-tą a jej rodzice j-tą

Chcemy znaleźć sieć G maksymalizującą powyższą miarę.

Uczenie struktury sieci bayesowskiej – miara jakości

Dane *d*:

X_1	X_2
1	1
1	2
1	1
2	2
1	1
2	1
1	1
2	2

*G*1

$P(X_1 = 1) = 7/12$

$P(X_2 = 1 \mid X_1 = 1) = 5/7$

$P(X_2 = 1 \mid X_1 = 2) = 2/5$

$$P(d \mid G1) = \left(\frac{\Gamma(4)}{\Gamma(4+8)}\frac{\Gamma(2+5)\Gamma(2+3)}{\Gamma(2)\Gamma(2)}\right)\left(\frac{\Gamma(2)}{\Gamma(2+5)}\frac{\Gamma(1+4)\Gamma(1+1)}{\Gamma(1)\Gamma(1)}\right)\left(\frac{\Gamma(2)}{\Gamma(2+3)}\frac{\Gamma(1+1)\Gamma(1+2)}{\Gamma(1)\Gamma(1)}\right) = 7.215\times10^{-6}$$

$$P(G1 \mid d) = \frac{P(d \mid G1)P(G1)}{P(d)}$$

*G*2

$$P(d \mid G2) = 6.746\times10^{-6}$$

$$P(G2 \mid d) = \frac{P(d \mid G2)P(G2)}{P(d)}$$

Uczenie struktury sieci, metody przybliżone

- Przeszukiwana jest przestrzeń zawierająca wszystkie kandydujące rozwiązania.
- Definiuje się zbiór operacji przekształcających jedno rozwiązanie w drugie.

Uczenie struktury sieci, metody przybliżone

Lokalna ocena sieci w węźle X_i:

$$score_B(d, X_i, PA_i) = \prod_{j=1}^{q_i^{(PA)}} \frac{\Gamma(N_{ij}^{(PA)})}{\Gamma(N_{ij}^{(PA)} + M_{ij}^{(PA)})} \prod_{k=1}^{r_i} \frac{\Gamma(a_{ijk}^{(PA)} + s_{ijk}^{(PA)})}{\Gamma(a_{ijk}^{(PA)})}$$

$$\Downarrow$$

$$score_B(d, G) = \prod_{i=1}^{n} score_B(d, X_i, PA_i)$$

Uczenie sieci, metody przybliżone - algorytm K2

Założenie: dany jest porządek węzłów (jeśli X_i jest przed X_j to nie jest dozwolona krawędź $X_j \rightarrow X_i$).

```
for i = 1 to n                    {dla każdego węzła}
  PA_i = ∅                        {PA_i - rodzice węzła i-tego}
  P = score_B(d, X_i, PA_i)
  repeat
    znajdź Y maksymalizujący score_B(d, X_i, PA_i ∪ {Y})
    P' = score_B(d, X_i, PA_i ∪ {Y})
    if P' > P then
      P = P'
      PA_i = PA_i ∪ {Y}           {dodanie nowego rodzica}
    end
  until nie znaleziono węzła Y, dla którego P' > P
end
```

Uczenie sieci, metody przybliżone - algorytm nie wymagający uporządkowania węzłów

Założenie: dozwolone są następujące operacje: dodanie krawędzi między dwoma węzłami, usunięcie krawędzi, zmiana kierunku krawędzi (pod warunkiem, że nie powstaje cykl)

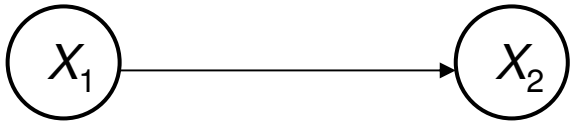
```
repeat
  if w sąsiedztwie sieci G istnieje sieć o większym score_B(d,G)
  then zastąp G siecią o największej wartości score_B(d,G)
  end
until żadna z operacji nie zwiększa wartości score_B(d,G)
```

Sąsiedztwo sieci G – wszystkie sieci, które można otrzymać z sieci G wykonując jedną z dozwolonych operacji.

Problem: możliwość znalezienia lokalnego maksimum

Równoważność sieci

- $\rho(f_{11}|d) = beta(f_{11}; 2 + 4, 2 + 3)$
- $\rho(f_{21}|d) = beta(f_{21}; 1 + 3, 1 + 1)$
- $\rho(f_{22}|d) = beta(f_{22}; 1 + 2, 1 + 1)$



$P(X_1=1)=6/11$

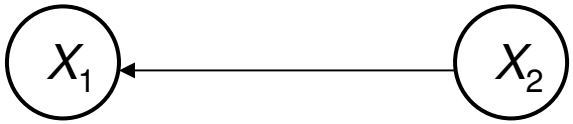
$P(X_2=1 \mid X_1=1)=2/3$
$P(X_2=1 \mid X_1=2)=3/5$

Dane d:

X_1	X_2
1	1
1	1
1	1
1	2
2	1
2	1
2	2

$\boldsymbol{P(X_1=1,X_2=1)} = P(X_2=1 \mid X_1=1)\ P(X_1=1) =$
$= 2/3 * 6/11 = \mathbf{4/11}$
$\boldsymbol{P(X_1=1,X_2=2)} = P(X_2=2 \mid X_1=1)\ P(X_1=1) =$
$= 1/3 * 6/11 = \mathbf{2/11}$
$\boldsymbol{P(X_1=2,X_2=1)} = P(X_2=1 \mid X_1=2)\ P(X_1=2) =$
$= 3/5 * 5/11 = \mathbf{3/11}$
$\boldsymbol{P(X_1=2,X_2=1)} = \mathbf{2/11}$

- $\rho(f_{22}|d) = beta(f_{11}; 2 + 5, 2 + 2)$
- $\rho(f_{11}|d) = beta(f_{21}; 1 + 3, 1 + 2)$
- $\rho(f_{12}|d) = beta(f_{22}; 1 + 1, 1 + 1)$



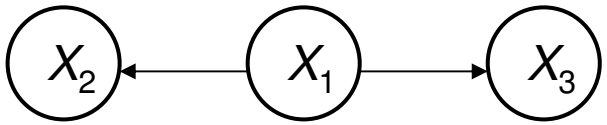
$P(X_1=1 \mid X_2=1)=4/7$
$P(X_1=1 \mid X_2=2)=1/2$

$P(X_2=1)=7/11$

$\boldsymbol{P(X_1=1,X_2=1)} = P(X_1=1 \mid X_2=1)\ P(X_2=1) =$
$= 4/7 * 7/11 = \mathbf{4/11}$
$\boldsymbol{P(X_1=1,X_2=2)} = P(X_1=1 \mid X_2=2)\ P(X_2=2) =$
$= 1/2 * 4/11 = \mathbf{2/11}$
$\boldsymbol{P(X_1=2,X_2=1)} = P(X_1=2 \mid X_2=1)\ P(X_2=1) =$
$= 3/7 * 7/11 = \mathbf{3/11}$
$\boldsymbol{P(X_1=2,X_2=1)} = \mathbf{2/11}$

jeden rozkład prawdopodobieństwa ↔ różne sieci

Równoważność sieci



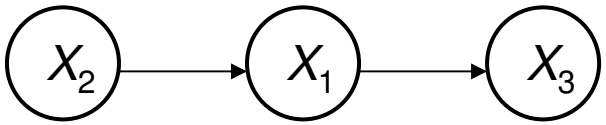
$P(X_2=1|X_1=1)=3/4$
$P(X_2=1|X_1=2)=2/3$

$P(X_3=1|X_1=1)=1/2$
$P(X_3=1|X_1=2)=1/3$

Zbiór uczący:

X_1	X_2	X_3
1	1	2
2	2	1
1	2	1
2	1	2
1	1	1
2	1	2
1	1	2

$P(X_1=1|X_2=1)=3/5$
$P(X_1=1|X_2=2)=1/2$



$P(X_2=1)=5/7$

$P(X_3=1|X_1=1)=1/2$
$P(X_3=1|X_1=2)=1/3$

$P(X_1=1,X_2=1,X_3=1)$ = ?

$P(X_1=1) * P(X_2=1|X_1=1) * P(X_3=1|X_1=1) =$
$= 4/7 * 3/4 * 1/2=$ **3/14**

$P(X_2=1) * P(X_1=1|X_2=1) * P(X_3=1|X_1=1) =$
$= 5/7 * 3/5 * 1/2 =$ **3/14**

$P(X_1=2,X_2=2,X_3=2)$ = ?

$P(X_1=2) * P(X_2=2|X_1=2) * P(X_3=2|X_1=2) =$
$= 3/7 * 1/3 * 2/3=$ **2/21**

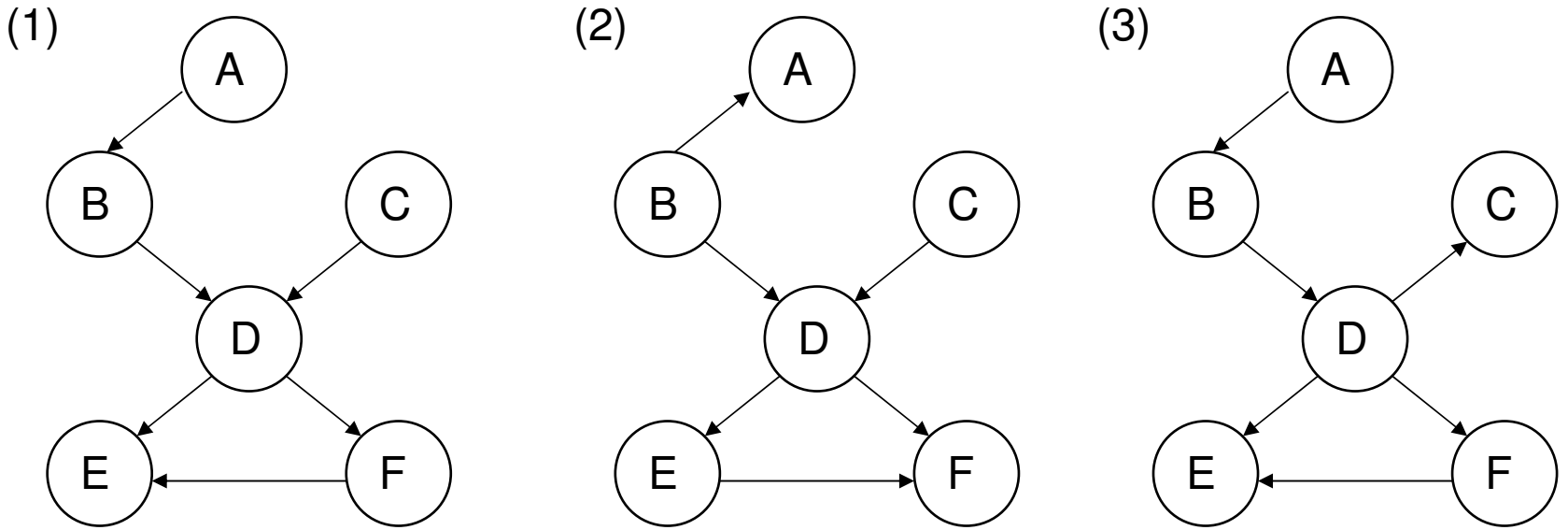
$P(X_2=2) * P(X_1=2|X_2=2) * P(X_3=2|X_1=2) =$
$= 2/7 * 1/2 * 2/3 =$ **2/21**

.....

jeden rozkład prawdopodobieństwa ↔ różne sieci

Równoważność sieci

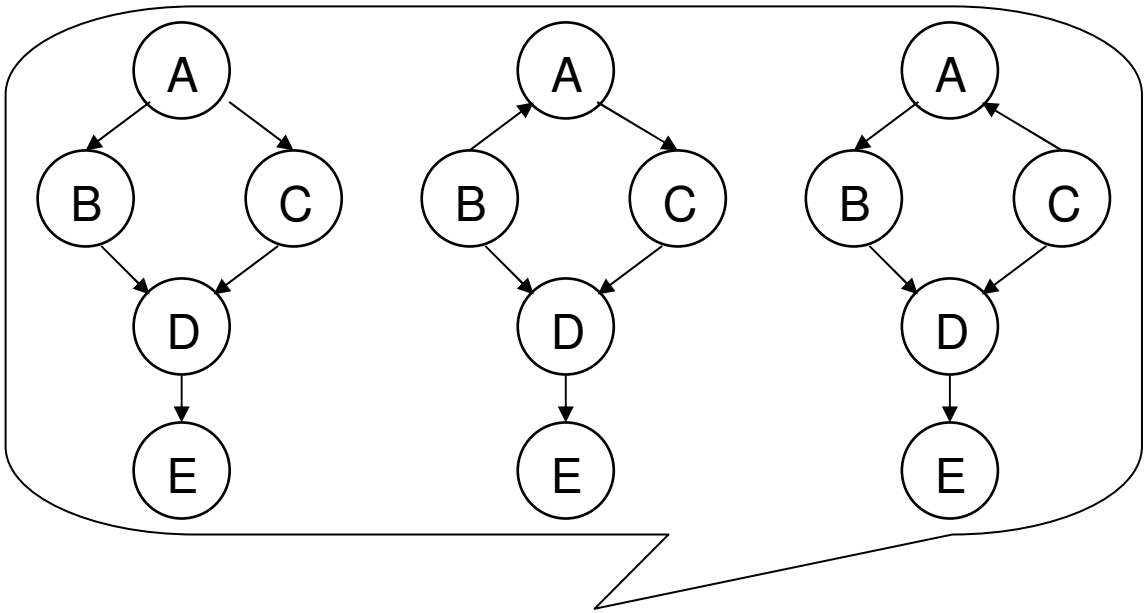
Twierdzenie: dwie sieci zawierające ten sam zbiór zmiennych są równoważne, gdy mają te same krawędzie (niezależnie od ich kierunku) oraz takie same połączenia typu X→Y←Z w przypadku, gdy nie ma krawędzi między X i Z [Pearl 1989].



(1) i (2) są równoważne
(1) i (3) nie są równoważne
(2) i (3) nie są równoważne

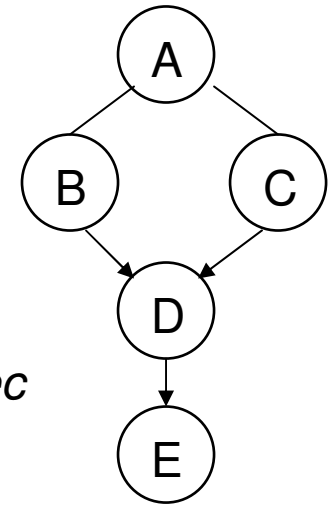
Równoważność sieci - wzorzec

Wzorzec klasy sieci równoważnych – sieć, która ma takie same połączenia jak sieci w tej klasie i tylko połączenia wspólne dla wszystkich sieci tej klasy są skierowane.



klasa sieci równoważnych

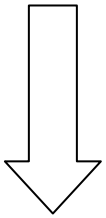
wzorzec



Poszukiwanie sieci czy wzorca?

- Jeśli w klasach sieci równoważnych jest dużo sieci, to znacznie dłuższy czas poszukiwania (strata czasu, ponieważ wszystko jedno którą sieć z klasy sieci równoważnych wybierzemy).

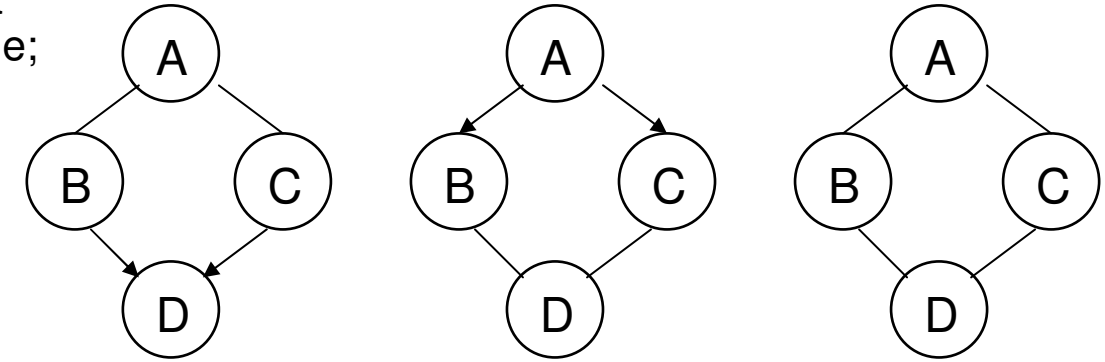
- Traktowanie wszystkich sieci jak jednakowo prawdopodobnych (jeśli w klasie jest dużo sieci, to duże szanse na wybór jednej z nich $\rightarrow$ różne rozkłady łączne nie są traktowane jednakowo).



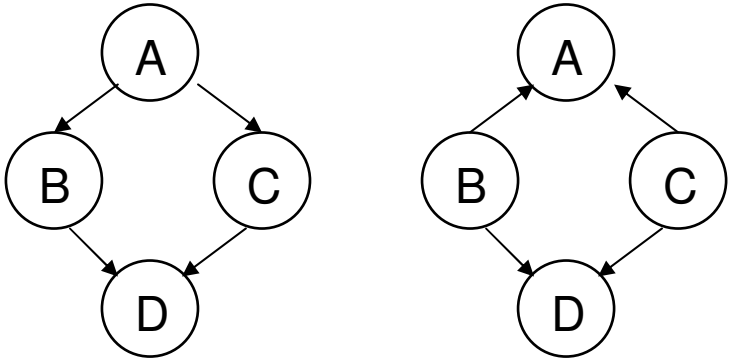
Algorytmy znajdujące najlepsze wzorce

Uczenie struktury sieci, metody przybliżone – algorytm znajdujący najlepszy wzorzec

- graf częściowo skierowany – zawiera krawędzie skierowane i nieskierowane;



- sieć *G* odpowiadająca grafowi częściowo skierowanemu *g* – wszystkie krawędzie skierowane w *g* są skierowane w *G*, *G* nie zawiera połączeń typu X→Y←Z (bez krawędzi między X a Z) innych niż w *g*;



- każdy wzorzec sieci jest grafem częściowo skierowanym, ale nie każdy graf częściowo skierowany jest wzorcem;

Uczenie struktury sieci, metody przybliżone – algorytm znajdujący najlepszy wzorzec

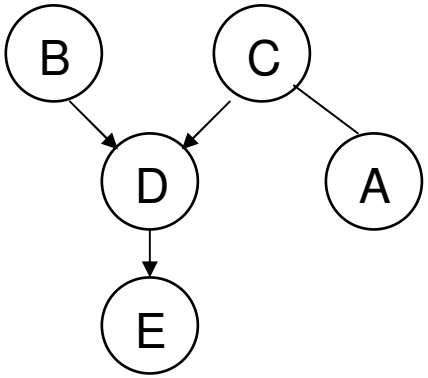
Zestaw operacji przekształcających graf częściowo skierowany *g1* w *g2*:

- dodanie krawędzi (dowolnie skierowanej lub nieskierowanej) między dwoma węzłami
- usunięcie krawędzi nieskierowanej
- usunięcie lub zmiana kierunku krawędzi skierowanej
- zamiana połączenia typu X-Y-Z (nie ma krawędzi między X a Z) na X→Y←Z

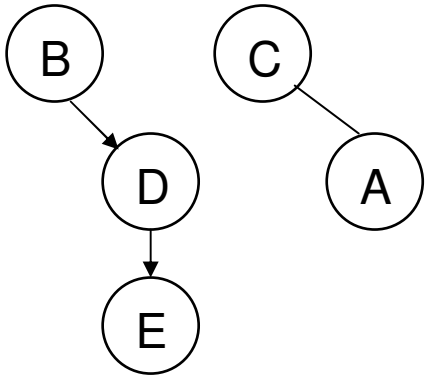
Przekształcanie jednego wzorca w drugi:

```
wykonaj jedną z operacji zmieniając wzorzec g na g'
    {g' może nie być wzorcem}
znajdź sieć G odpowiadającą grafowi g'
if znaleziono sieć G then
  znajdź wzorzec g'' odpowiadający sieci G
end if
```

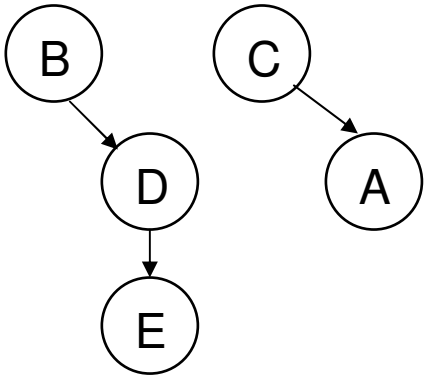
Uczenie struktury sieci, metody przybliżone – algorytm znajdujący najlepszy wzorzec



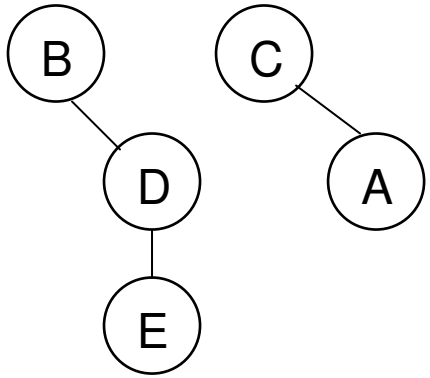
1) usunięcie krawędzi



2) znalezienie sieci



3) znalezienie wzorca



Uczenie struktury sieci, metody przybliżone – algorytm znajdujący najlepszy wzorzec

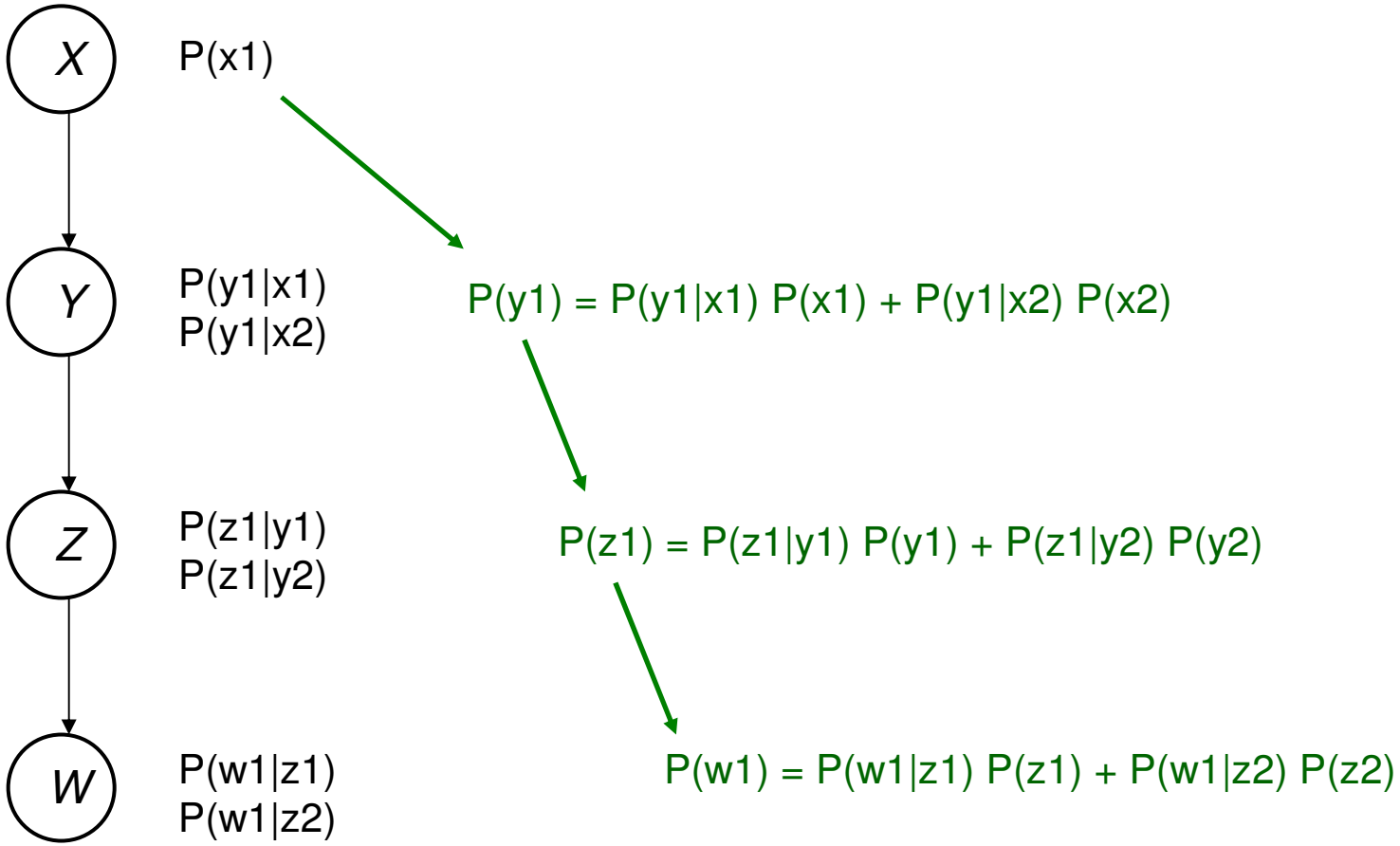
```
repeat
  if w sąsiedztwie wzorca gp istnieje wzorzec o większym
    score_B(d,gp)
  then zastąp dany wzorzec wzorcem poprawiającym score_B(d,gp)
    najbardziej
  end
until żadna z operacji nie zwiększa wartości score_B(d,G)
```

W przeciwieństwie do dwóch poprzednich algorytmów, tym razem nie można uaktualniać miary $score_B$ lokalnie.

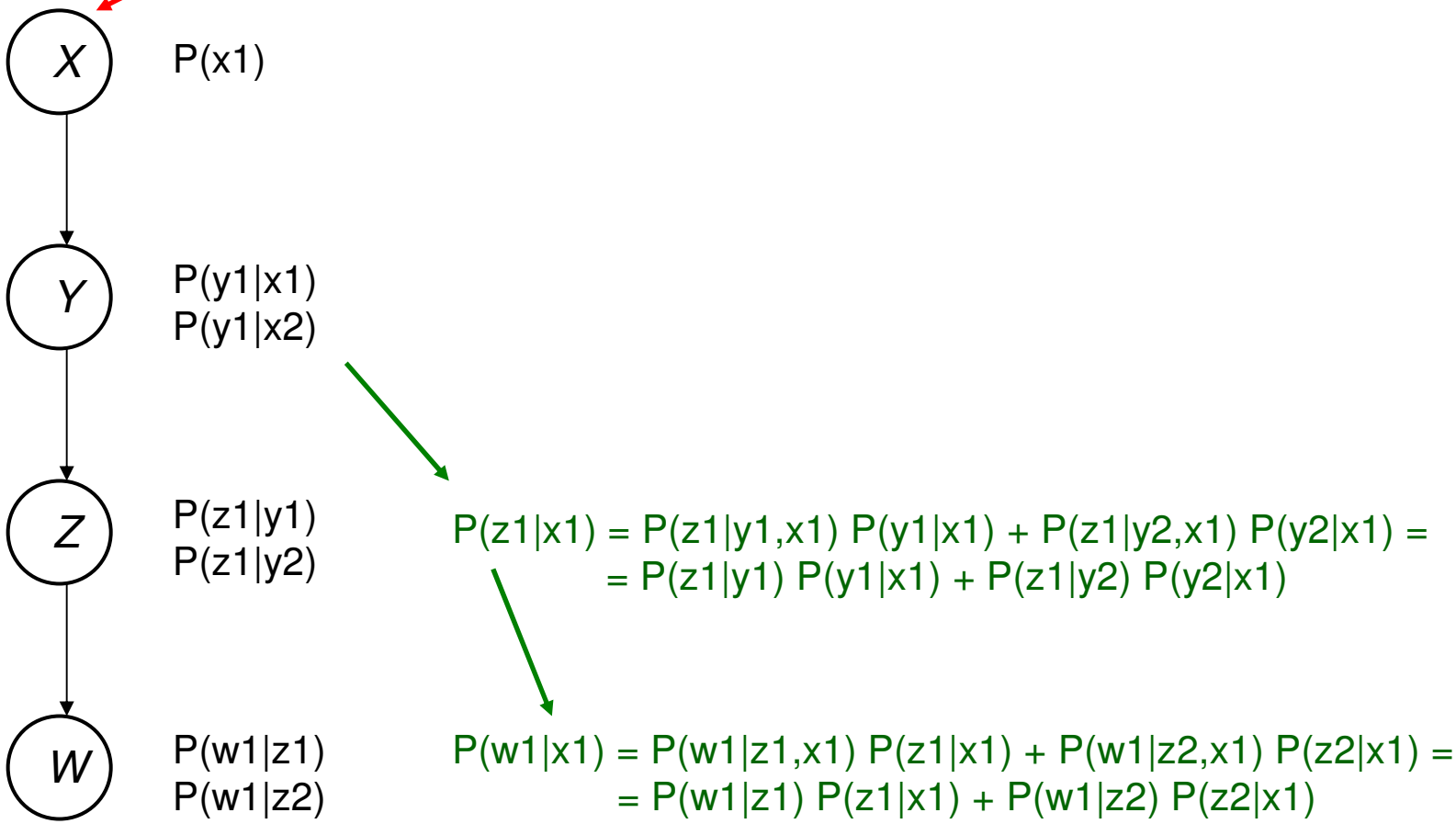
Uczenie struktury sieci, metody przybliżone - porównanie

- Metody poszukujące wzorca przeważnie prowadzą do lepszych sieci niż metody znajdujące sieć.
- Metody poszukujące wzorca są wolniejsze.
- Jeżeli znane jest uporządkowanie zmiennych wymagane przez algorytm K2, to algorytm ten znajduje sieci często lepsze niż inne metody.

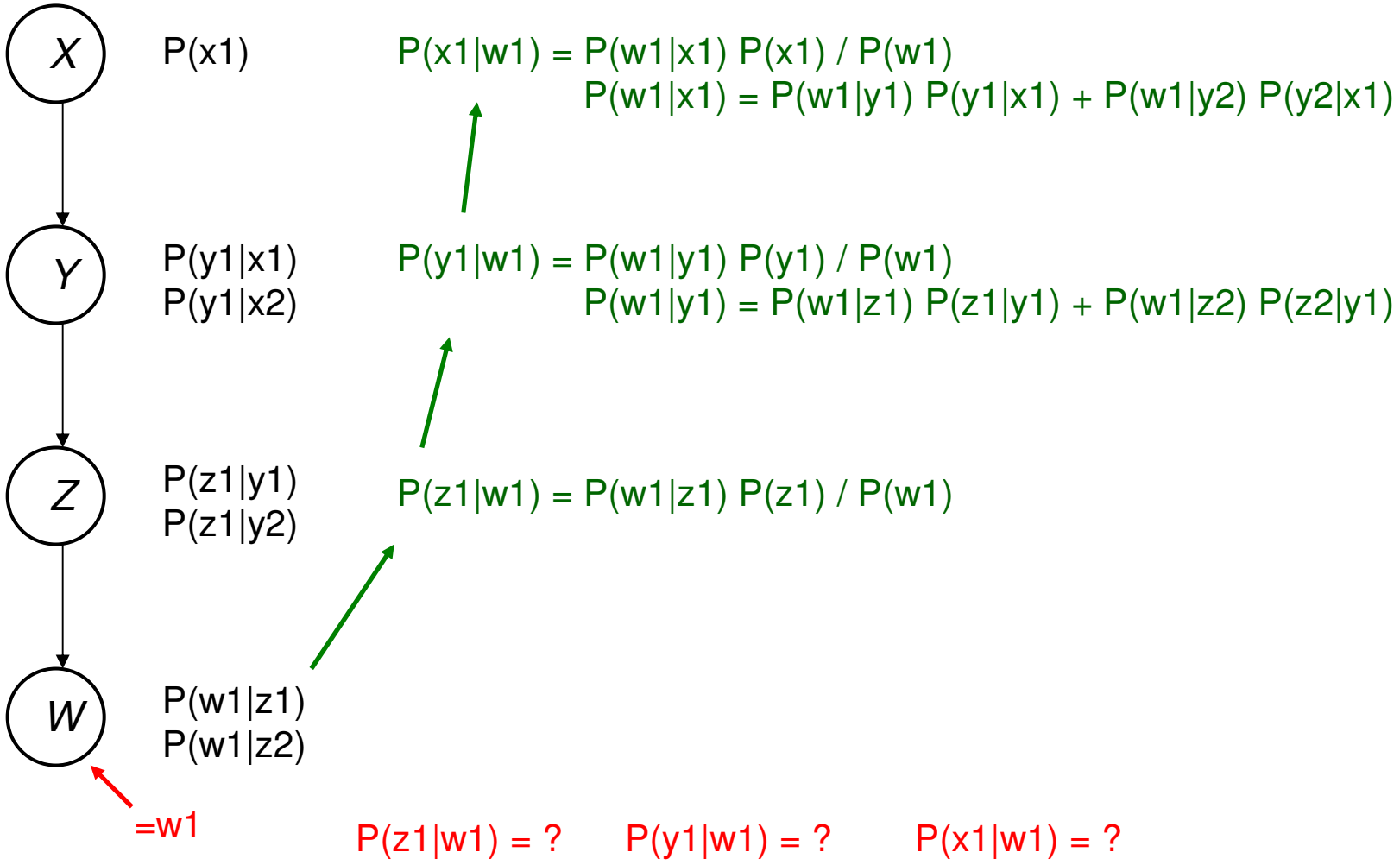
Wnioskowanie



Wnioskowanie



Wnioskowanie



Wnioskowanie