# Algorytmy ewolucyjne w praktyce

algorytmy Estimation of Distribution Algorithms oparte na rozkładach prawdopodobieństwa modelowanych przez sieci bayesowskie i gaussowskie

Piotr Lipiński

## Estimation of Distribution Algorithms

- Proste algorytmy EDA oparte na rozkładach prawdopodobieństwa wektora niezależnych zmiennych losowych:
  - dla dyskretnych problemów optymalizacji
    - UMDA
    - PBIL
    - CGA
  - dla ciągłych problemów optymalizacji
    - UMDAc
    - SHCLVND
    - PBILc
- Bardziej zaawansowane algorytmy EDA oparte na rozkładach prawdopodobieństwa wektora skorelowanych zmiennych losowych:
  - dla dyskretnych problemów optymalizacji
    - MIMIC, COMIT, EBNA, BOA, hBOA
  - dla ciągłych problemów optymalizacji
    - EGNA, modyfikacje powyższych

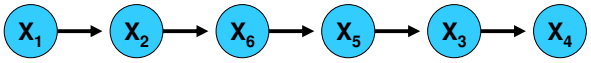
## Estimation of Distribution Algorithms

- Popularnym podejściem do modelowania zależności między zmiennymi losowymi w chromosomie jest użycie probabilistycznych modeli graficznych.
- Probabilistyczne modele graficzne używają grafów do reprezentowania łącznych rozkładów prawdopodobieństwa modelujących wielowymiarowe wektory losowe.
- Najważniejsze probabilistyczne modele graficzne to:
  - sieci bayesowskie (ang. *Bayesian Networks*),
  - losowe pola Markowa (ang. *Markov Random Fields*).



## Mutual Information Maximization for Input Clustering (MIMIC)

- MIMIC modeluje zależności między zmiennymi losowymi za pomocą łańcucha (uproszczonej sieci bayesowskiej), tzn. zakłada, że zmienne losowe zależą kolejno od siebie, niekoniecznie jednak w kolejności wyznaczonej przez ich pozycje w chromosomie.
- Przykład:




- Prowadzi to do modelu probabilistycznego opartego na warunkowych rozkładach prawdopodobieństwa

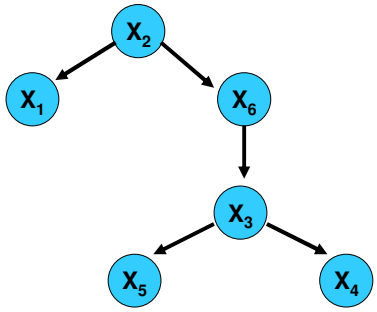
$$p(X) = p(X_{i_1} \mid X_{i_2})\, p(X_{i_2} \mid X_{i_3}) \ldots p(X_{i_{d-1}} \mid X_{i_d})\, p(X_{i_d})$$

gdzie $(i_1, i_2, \ldots, i_d)$ to pewna permutacja zbioru $\{1, 2, \ldots, d\}$.

- Kolejność zmiennych losowych w łańcuchu kodowana jest więc za pomocą permutacji ustalanej na podstawie testów statystycznych zależności zmiennych losowych opartych na bieżącej populacji.
- Prawdopodobieństwa warunkowe są estymowane z próbki danych tworzonej przez bieżącą populację.

## Combining Optimizers with Mutual Information Trees (COMIT)

- COMIT rozszerza podejście stosowane w algorytmie MIMIC. Zamiast łańcucha, używa drzewa modelującego zależności między zmiennymi losowymi.
- Przykład:




modelowany rozkład prawdopodobieństwa to

$$p(X) = p(X_2)\, p(X_1 \mid X_2)\, p(X_6 \mid X_2)\, p(X_3 \mid X_6)\, p(X_5 \mid X_3)\, p(X_4 \mid X_3)$$



## Combining Optimizers with Mutual Information Trees (COMIT)

- Podobnie jak w MIMIC, struktura drzewa ustalana jest na podstawie testów statystycznych zależności zmiennych losowych opartych na bieżącej populacji.
- Prawdopodobieństwa warunkowe są estymowane z próbki danych tworzonej przez bieżącą populację.

## Sieci bayesowskie

- Sieci bayesowskie opierają się na bayesowskim podejściu do prawdopodobieństwa i statystyki:
  - prawdopodobieństwo bayesowskie: stopień przekonania o wystąpieniu określonego zdarzenia
  - prawdopodobieństwo klasyczne: prawdziwe (teoretyczne) prawdopodobieństwo wystąpienia określonego zdarzenia

- Twierdzenie Bayesa: prawdopodobieństwo *a posteriori* wystąpienia zdarzenia A pod warunkiem wystąpienia zdarzenia D i wiedzy $\xi$:

$$p(A \mid D, \xi) = \frac{p(A \mid \xi) p(D \mid A, \xi)}{p(D \mid \xi)}$$

  - Interpretacja:
    - p(A | D, ξ) to prawdopodobieństwo tego, że zaobserwowalibyśmy zdarzenie A, jeśli wystąpiłoby zdarzenie D i prawdziwa byłaby wiedza ξ,
    - p(D | A, ξ) to prawdopodobieństwo tego, że rzeczywiście wystąpiło zdarzenie D, jeśli zaobserwowaliśmy zdarzenie A i prawdziwa była wiedza ξ,
    - p(A | ξ) i p(D | ξ) to prawdopodobieństwo wystąpienia zdarzenia A i D, odpowiednio, przy założeniu, że prawdziwa jest wiedza ξ.

- Prowadzi to do funkcji wiarygodności:

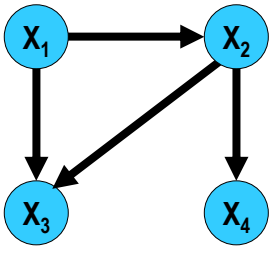
$$L(A; D) = p(D \mid A)$$



## Sieci bayesowskie

- Sieć bayesowska jest najpopularniejszym graficznym modelem probabilistycznym dla rozkładów dyskretnych. Jej odpowiednikiem dla rozkładów ciągłych jest sieć gaussowska.

- Sieć bayesowska to acykliczny graf skierowany, w którym wierzchołki reprezentują zmienne losowe, a krawędzie odpowiadają zależnościom między tymi zmiennymi.
  - Zwrot krawędzi reprezentuje "kierunek patrzenia" na zależności między zmiennymi losowymi (zmienna w węźle końcowym zależy od tej w węźle początkowym). Brak krawędzi między dwoma wierzchołkami oznacza, że zmienne losowe reprezentowane przez te wierzchołki są niezależne.



## Sieci bayesowskie

- Sieć bayesowska modeluje rozkład prawdopodobieństwa w następujący sposób:
  - Niech $\mathbf{\Pi}_{X_i}$ oznacza zbiór "rodziców" wierzchołka $X_i$.
  - Wartość zmiennej losowej $X_i$ można więc określić warunkowym rozkładem prawdopodobieństwa $P(X_i \mid \mathbf{\Pi}_{X_i})$.
  - Wartość całego wektora losowego $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ można więc określić łącznym rozkładem prawdopodobieństwa

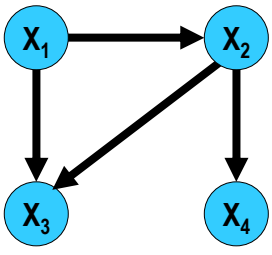
$$\prod_{i=1}^{n} P(X_i \mid \Pi_{X_i})$$



## Sieci bayesowskie w EDA

- W algorytmach EDA, m.in. BOA i EBNA, sieci bayesowskie są używane do reprezentowania zależności między genami w chromosomie.

- Przykład:
  - Rozpatrzmy chromosom X o długości d = 4.
  - Wartość każdego genu $X_1$, $X_2$, $X_3$, $X_4$ można traktować jako wartość pewnej zmiennej losowej. Wartość całego chromosomu można traktować jako wartość pewnego wektora losowego $\mathbf{X} = (X_1, X_2, X_3, X_4)$.
  - Poniższa sieć bayesowska może więc reprezentować zależności między genami w takim chromosomie.



## Sieci bayesowskie w EDA

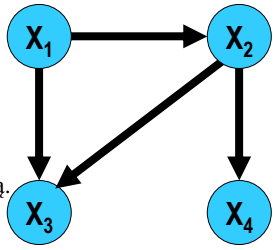
- Interpretacja takiej sieci bayesowskiej jest następująca:
  - wartość genu $X_1$ jest niezależna od wartości pozostałych genów,
  - wartość genu $X_2$ zależy od wartości genu $X_1$,
  - wartość genu $X_3$ zależy od wartości genu $X_1$ i wartości genu $X_2$,
  - wartość genu $X_4$ zależy od wartości genu $X_2$.

- Łączny rozkład prawdopodobieństwa wektora losowego $\mathbf{X} = (X_1, X_2, X_3, X_4)$ można więc przedstawić jako

$$P(\mathbf{X} = \mathbf{x}) = P(X_1 = x_1)\, P(X_2 = x_2 \mid X_1 = x_1)\, P(X_3 = x_3 \mid X_1 = x_1, X_2 = x_2)\, P(X_4 = x_4 \mid X_2 = x_2)$$

dla $\mathbf{x} = (x_1, x_2, x_3, x_4)$.

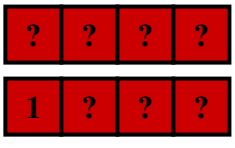
- Ten wzór można zastosować przy generowaniu nowych osobników według modelu zadanego przez sieć bayesowską.

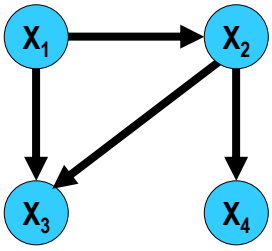





## Sieci bayesowskie w EDA

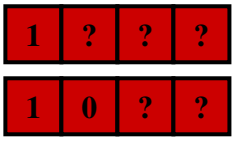
- Generowanie osobnika na podstawie modelu probabilistycznego opisywanego przez sieć bayesowską:
  - Najpierw losujemy wartości zmiennych niezależnych.
  - Znając prawdopodobieństwa $P(X_1 = 0)$ i $P(X_1 = 1)$ z jakimi zmienna niezależna $X_1$ przyjmuje odpowiednio wartości 0 i 1, losujemy wartość genu $X_1$.
  - Jeśli byłyby inne zmienne niezależne, postępujemy z nimi w taki sam sposób. Kolejność rozpatrywania tych zmiennych jest nieistotna.

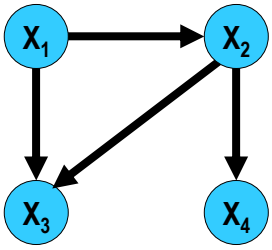


## Sieci bayesowskie w EDA

- W kolejnych krokach zajmujemy się zmiennymi, których rodzice mają już ustalone wartości. Kolejność rozpatrywania tych zmiennych jest nieistotna.
- Znając prawdopodobieństwa $P(X_2 = 0 \mid X_1 = 1)$ i $P(X_2 = 1 \mid X_1 = 1)$, losujemy wartość genu $X_2$.

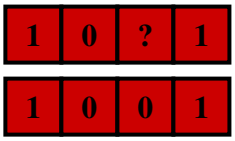


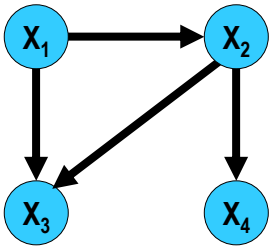






## Sieci bayesowskie w EDA

- Mając ustalone wartości genów $X_1$ i $X_2$, możemy losować wartości genów $X_3$ i $X_4$.
- Zajmijmy się najpierw zmienną $X_4$. Znając prawdopodobieństwa $P(X_4 = 0 \mid X_2 = 0)$ i $P(X_4 = 1 \mid X_2 = 0)$, losujemy wartość genu $X_4$.
- Na koniec zajmijmy się zmienną $X_3$. Znając prawdopodobieństwa $P(X_3 = 0 \mid X_1 = 1, X_2 = 0)$ i $P(X_3 = 1 \mid X_1 = 1, X_2 = 0)$, losujemy wartość genu $X_3$.

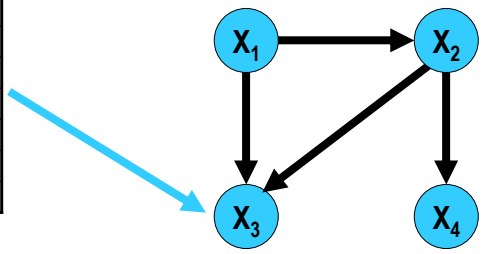


## Sieci bayesowskie w EDA

- Z każdym węzłem sieci bayesowskiej jest związany rozkład prawdopodobieństwa, reprezentowany przez tabelę prawdopodobieństw warunkowych.
- W przypadku rozkładów ciągłych (sieci gaussowskich) zamiast tabeli prawdopodobieństw warunkowych podawana jest gęstość rozkładu prawdopodobieństwa.

| $X_3$ | $X_1$ | $X_2$ | $P(X_3 \mid X_1, X_2)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.25 |
| 0 | 0 | 1 | 0.10 |
| 0 | 1 | 0 | 0.50 |
| 0 | 1 | 1 | 0.95 |
| 1 | 0 | 0 | 0.75 |
| 1 | 0 | 1 | 0.90 |
| 1 | 1 | 0 | 0.50 |
| 1 | 1 | 1 | 0.05 |






## Bayesian Optimization Algorithm (BOA)

- Algorytm BOA:

```
P = Random-Population(N);
Population-Evaluation(P, F);
B = Bayesian-Network-Initialization();
while not Termination-Condition()
    P^(S) = Select(P, N, M);
    B = Bayesian-Network-Update(B, P^(S));
    P = Random-Population(B, N);
    Population-Evaluation(P, F);
return best of P;
```

- Znaczenie parametrów:
  - F – funkcja celu
  - N – liczba osobników w populacji głównej **P**
  - M – liczba osobników wybieranych z populacji głównej **P** do populacji $\mathbf{P}^{(S)}$ (wielkość próbki)

## Bayesian Optimization Algorithm (BOA)

- Wyjaśnienia wymaga sposób tworzenia sieci bayesowskiej oraz estymacji prawdopodobieństw warunkowych używanych m.in. przy generowaniu nowej populacji losowych osobników według modelu zadanego przez sieć bayesowską.
  - Na początku algorytmu sieć bayesowska **B** jest siecią pustą (bez krawędzi).
  - W kolejnych iteracjach algorytm modyfikuje sieć bayesowską **B** na podstawie próbki danych $\mathbf{P}^{(S)}$ zawierającej najlepsze osobniki z aktualnej populacji **P**.
  - Wraz z modyfikacją sieci bayesowskiej algorytm estymuje prawdopodobieństwa warunkowe (przechowywane w tabelach związanych z węzłami sieci bayesowskiej) na podstawie próbki danych $\mathbf{P}^{(S)}$ zawierającej najlepsze osobniki z aktualnej populacji **P**.
- Do oceny dopasowania sieci bayesowskiej do próbki danych używa się bayesowskiej metryki Dirichleta lub jej uproszczonej wersji – metryki K2.



## Bayesian Dirichlet Metric

Dopasowanie sieci bayesowskiej do próbki danych ocenia się różnymi miarami, zazwyczaj opartymi na funkcji wiarygodności. Jedną z najpopularniejszych miar jest bayesowska metryka Dirichleta (ang. *Bayesian Dirichlet Metric*):

$$p(\mathbf{B}, \mathbf{P}^{(S)} \mid \xi) = p(\mathbf{B} \mid \xi) \prod_{i=1}^{d} \prod_{\pi_{X_i}} \frac{m'(\pi_{X_i})!}{(m'(\pi_{X_i}) + m(\pi_{X_i}))!} \prod_{x_i} \frac{(m'(x_i, \pi_{X_i}) + m(x_i, \pi_{X_i}))!}{m'(x_i, \pi_{X_i})!}$$

gdzie

iloczyn po $\pi_{xi}$ jest po wszystkich konfiguracjach rodziców wierzchołka $X_i$
iloczyn po $x_i$ jest po wszystkich wartościach wierzchołka $X_i$

| | |
|---|---|
| **B** | sieć bayesowska |
| $\mathbf{P}^{(S)}$ | próbka danych zawierająca najlepsze osobniki z bieżącej populacji |
| $\xi$ | ewentualna dodatkowa wiedza *a priori* |
| $m(\pi_{xi})$ | liczba osobników w $\mathbf{P}^{(S)}$, w których rodzice wierzchołka $X_i$ mają konfigurację $\pi_{xi}$ |
| $m(x_i, \pi_{xi})$ | liczba osobników w $\mathbf{P}^{(S)}$, w których wierzchołek $X_i$ ma wartość $x_i$, a jego rodzice mają konfigurację $\pi_{xi}$ |
| $p(\mathbf{B}\|\xi)$ | dodatkowy współczynnik prawdopodobieństwa *a priori* (zazwyczaj $p(\mathbf{B}\|\xi) = 1$) |
| $m'(\ldots)$ | dodatkowa informacja *a priori*, analogiczna do $m(\ldots)$ (zazwyczaj $m(x_i, \pi_{xi}) = 1$, a $m(\pi_{xi}) = m(0, \pi_{xi}) + m(1, \pi_{xi}) = 2$) |

# Metryka K2

Jeśli nie wykorzystujemy wiedzy *a priori*, bayesowska metryka Dirichleta upraszcza się do metryki K2:

$$p(\mathbf{B},\mathbf{P}^{(S)}) = \prod_{i=1}^{d}\prod_{\pi_{x_i}}\frac{2!}{(2+m(\pi_{X_i}))!}\prod_{x_i}(1+m(x_i,\pi_{X_i}))!$$

gdzie

iloczyn po $\pi_{xi}$ jest po wszystkich konfiguracjach rodziców wierzchołka $X_i$
iloczyn po $x_i$ jest po wszystkich wartościach wierzchołka $X_i$

| | |
|---|---|
| **B** | sieć bayesowska |
| $\mathbf{P}^{(S)}$ | próbka danych zawierająca najlepsze osobniki z bieżącej populacji |
| $m(\pi_{xi})$ | liczba osobników w $\mathbf{P}^{(S)}$, w których rodzice wierzchołka $X_i$ mają konfigurację $\pi_{xi}$ |
| $m(x_i, \pi_{xi})$ | liczba osobników w $\mathbf{P}^{(S)}$, w których wierzchołek $X_i$ ma wartość $x_i$, a jego rodzice mają konfigurację $\pi_{xi}$ |



# Metryka K2 – przykład

- Rozpatrzmy dwie zmienne losowe $X_1$ i $X_2$ tworzące chromosom oraz populację

  $\mathbf{P}^{(S)}$ = { 00, 00, 00, 11 }

- Policzmy K2 dla sieci bayesowskiej $\mathbf{B}_{empty}$ bez żadnych krawędzi:

  Ponieważ $m'(x_i, \pi_{xi}) = 1$, to $m'(\pi_{xi}) = m'(0, \pi_{xi}) + m'(1, \pi_{xi}) = 2$.

  Łatwo wyliczyć, że

| i | $x_i$ | $\pi_{x_i}$ | $m(x_i, \pi_{x_i})$ |
|---|---|---|---|
| 1 | 0 | 0 | 3 |
| 1 | 1 | 0 | 1 |
| 2 | 0 | 0 | 3 |
| 2 | 1 | 0 | 1 |

  Zatem

$$p(\mathbf{B}_{empty},\mathbf{P}^{(S)}) = \frac{2!}{(2+4)!}(1+3)!(1+1)!\frac{2!}{(2+4)!}(1+3)!(1+1)! = \frac{4}{225}$$

## Metryka K2 – przykład

- Policzmy K2 dla sieci bayesowskiej $\mathbf{B}_{1->2}$ z krawędzią od $X_1$ do $X_2$:

  Ponieważ $m'(x_i, \pi_{xi}) = 1$, to $m'(\pi_{xi}) = m'(0, \pi_{xi}) + m'(1, \pi_{xi}) = 2$.

  Łatwo wyliczyć, że

| i | $x_i$ | $\pi_{x_i}$ | $m(x_i, \pi_{x_i})$ |
|---|---|---|---|
| 1 | 0 | 0 | 3 |
| 1 | 1 | 0 | 1 |
| 2 | 0 | 0 | 3 |
| 2 | 0 | 1 | 0 |
| 2 | 1 | 0 | 0 |
| 2 | 1 | 0 | 1 |

  Zatem

$$p(\mathbf{B}_{1->2}, \mathbf{P}^{(S)}) = \frac{2!}{(2+4)!}(1+3)!(1+1)!\frac{2!}{(2+3)!}(1+3)!(1+0)!\frac{2!}{(2+1)!}(1+1)!(1+0)! = \frac{8}{225}$$

- Sieć bayesowska $\mathbf{B}_{1->2}$ zatem lepiej modeluje próbkę danych niż sieć $\mathbf{B}_{empty}$.



## Bayesian Optimization Algorithm (BOA)

- Jak skonstruować sieć bayesowską **B** dobrze modelującą próbkę danych $\mathbf{P}^{(S)}$?
  - Naiwne podejście to sprawdzenie wszystkich możliwych sieci bayesowskich i wybranie sieci najlepiej modelującej próbkę danych.
    - W praktyce liczba wszystkich możliwych sieci bayesowskich jest zbyt duża do przejrzenia.
    - Można wprowadzić ograniczenia na rozpatrywane sieci bayesowskie, na przykład przez ograniczenie maksymalnego stopnia wierzchołka sieci bayesowskiej, co przy niewielkich problemach optymalizacji pozwoli przejrzeć wszystkie sieci bayesowskie.
  - Bardziej efektywne jest podejście heurystyczne, które zastosowano w algorytmie BOA.
    - Algorytm rozpoczyna działanie z pustą siecią bayesowską bez krawędzi.
    - W każdej iteracji algorytmu ewolucyjnego, algorytm modyfikuje sieć bayesowską **B** na podstawie próbki danych $\mathbf{P}^{(S)}$ używając algorytmu zachłannego, który próbuje poprawić aktualną sieć bayesowską przez wykonywanie następujących operacji:
      - dodanie losowej krawędzi,
      - usunięcie losowo wybranej krawędzi,
      - zmiana kierunku losowo wybranej krawędzi.

## Bayesian Optimization Algorithm (BOA)

- Jak estymować prawdopodobieństwa warunkowe (przechowywane w tabelach związanych z węzłami sieci bayesowskiej) na podstawie próbki danych $\mathbf{P}^{(S)}$?
  - Prawdopodobieństwa warunkowe są estymowane na podstawie próbki danych przez częstości występowania odpowiednich osobników w rozważanej próbce danych.
  - Przykład:

$$P( X_3 = 1 \mid X_1 = 0, X_2 = 0 ) = m / M$$

gdzie

M liczba osobników w $\mathbf{P}^{(S)}$, w których $X_1 = 0$ oraz $X_2 = 0$,

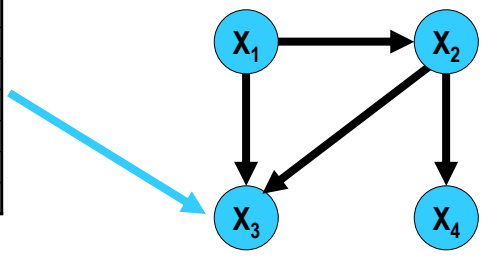
m liczba osobników w $\mathbf{P}^{(S)}$, w których $X_1 = 0$, $X_2 = 0$ oraz $X_3 = 1$.



## Hierarchical Bayesian Optimization Algorithm (hBOA)

- Jeden z problemów algorytmu BOA stanowią duże tabele prawdopodobieństw warunkowych związane z węzłami sieci bayesowskich.
  - Na przykład dla binarnej zmiennej losowej, która zależy od 10 innych binarnych zmiennych losowych, w tabeli musi znaleźć się $2 \times 2^{10} = 2^{11}$ wierszy!
  - Problem ten rozwiązano w algorytmie Hierarchical Bayesian Optimization Algorithm (hBOA), w którym wykorzystano drzewiaste struktury lokalne.

| $X_3$ | $X_1$ | $X_2$ | $P(X_3 \mid X_1, X_2)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.25 |
| 0 | 0 | 1 | 0.10 |
| 0 | 1 | 0 | 0.50 |
| 0 | 1 | 1 | 0.95 |
| 1 | 0 | 0 | 0.75 |
| 1 | 0 | 1 | 0.90 |
| 1 | 1 | 0 | 0.50 |
| 1 | 1 | 1 | 0.05 |



# Hierarchical Bayesian Optimization Algorithm (hBOA)

- Wykorzystanie struktur lokalnych w sieciach bayesowskich:
  - pełna tabela prawdopodobieństw warunkowych (jak w BOA),
  - tabela zawierająca "najważniejsze" prawdopodobieństwa warunkowe, pozostałe są aproksymowane na ich podstawie
    - Przykład: Tabela nie zawiera $P(X_3 = 0 \mid X_1 = 0, X_2 = 1)$, ale zawiera $P(X_3 = 0 \mid X_1 = 0)$. Przyjmuje się wówczas $P(X_3 = 0 \mid X_1 = 0, X_2 = 1) = P(X_3 = 0 \mid X_1 = 0)$,
  - wykorzystanie drzew decyzyjnych aproksymujących tabelę prawdopodobieństw warunkowych (jak w hBOA).

| $X_3$ | $X_1$ | $X_2$ | $P(X_3 \mid X_1, X_2)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.25 |
| 0 | 0 | 1 | 0.10 |
| 0 | 1 | 0 | 0.50 |
| 0 | 1 | 1 | 0.95 |
| 1 | 0 | 0 | 0.75 |
| 1 | 0 | 1 | 0.90 |
| 1 | 1 | 0 | 0.50 |
| 1 | 1 | 1 | 0.05 |