

Piotr Chmiel, 200608

Kamil Machnicki, 200752

Łukasz Matysiak, 200646

Grzegorz Zając, 200664

Jakub Zgraja, 200609

Miękkie metody obliczeniowe

DOKUMENTACJA PROJEKTOWA

- 1. SYSTEM ROZPOZNAWANIA OPARTY O SIECI BAYESOWSKIE**
- 2. SYSTEM ROZPOZNAWANIA OPARTY O UKRYTY MODEL
MARKOWA**

Rok akad. 2016/2017, kierunek INF, specjalność IMT, studia II stopnia

PROWADZĄCY:

dr inż. Konrad Jackowski

1 Wstęp teoretyczny

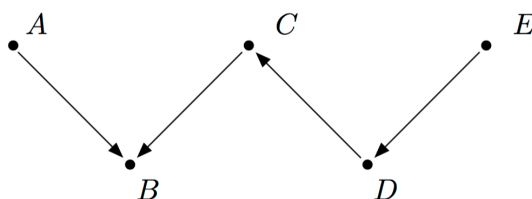
1.1 Sieci Bayesowskie

Definicja Sieci Bayesowskiej

Siecią Bayesa nazywamy skierowany graf acykliczny o wierzchołkach reprezentujących zmienne losowe i łukach określających zależności. Istnienie łuku pomiędzy dwoma wierzchołkami oznacza istnienie bezpośredniej zależności przyczynowo skutkowej pomiędzy odpowiadającymi im zmiennymi. Siła tej zależności określona jest przez tablice prawdopodobieństw warunkowych.

Inaczej mówiąc, sieć Bayesowska to acykliczny (nie zawierający cykli) graf skierowany, w którym:

- Węzły reprezentują zmienne losowe (np. temperaturę jakiegoś źródła, stan pacjenta, cechę obiektu itp.)
- Łuki (skierowane) reprezentują zależność typu „zmienna X ma bezpośredni wpływ na zmienną Y”,
- Każdy węzeł X ma stowarzyszona z nim tablice prawdopodobieństw warunkowych określających wpływ wywierany na X przez jego poprzedników (rodziców) w grafie.
- Zmienne reprezentowane przez węzły przyjmują wartości dyskretne.



Rysunek 1: Przykładowa sieć Bayesowska

Na rysunku został przedstawiony przykład nieskomplikowanej sieci Bayesowskiej. Na podstawie zaprezentowanego na rysunku grafu przedstawiającego sieć można wyciągnąć następujące wnioski:

- Para zmiennych A oraz B jest od siebie bezpośrednio zależna.
- Para zmiennych B oraz C jest od siebie bezpośrednio zależna.
- Zmienna reprezentowana przez wierzchołek B jest jednocześnie zależna od A oraz C.
- Zmienne A oraz C pozostają brzegowo niezależne do momentu ustalenia wartości B (własność tą można również określić mówiąc, że zmienne A i C są d-połączone).
- Zmienne E oraz C są zależne pośrednio.
- Zmienna D d-separuje zmienne C i E.
- Zmienne C i D d-separują parę zmiennych B, E.

Orientacja łuków jest konieczna do określenia zależności nieprzechodnych. Na przytoczonym przykładzie można zaobserwować, że pomimo faktu przemienności własności zależności orientacja krawędzi wnosi istotną informację o rozkładzie.

Sieć Bayesowska pozwala na wyznaczenie rozkładu prawdopodobieństwa zmiennych. Jeżeli przez $\Pi(X_i)$ oznaczmy zbiór rodziców danego wierzchołka w grafie, to rozkład prawdopodobieństwa zmiennych danej sieci opisuje się równaniem:

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^n P(X_i | \Pi(X_i))$$

Co dla przedstawionego powyżej przykładu wynosi:

$$P(A, B, C, D, E) = P(A)P(B|A, C)P(C|D)P(D|E)P(E)$$

Dzięki powyższemu równaniu równania możliwe jest określenie prawdopodobieństwa wystąpienia określonego wartościowania wszystkich zmiennych, znając jedynie lokalne prawdopodobieństwa warunkowe. Określając wartości tzw. *przyczyn podstawowych*, czyli węzłów grafu nie posiadających rodziców (zmienne A, E w podanym przykładzie), można określić wartości oczekiwane innych atrybutów. Wszystkie pozostałe zmienne zależą bowiem pośrednio lub bezpośrednio od tego zbioru.

Istotą działania sieci Bayesowskich jest propagacja informacji o rozkładzie prawdopodobieństwa. W praktyce jednak **sieci Bayesowskie używane są do ekstrakcji informacji o rozkładach nieznanych, o których wiadomości możemy czerpać tylko z dostępnych wyników prób z eksperymentów statystycznych**. Sieci te nie są zatem używane do opisywania dokładnych rozkładów.

Próby z eksperymentów statystycznych (zestawy danych uczących) mogą mieć bardzo duże rozmiary, a ekstrakcja zawartych w nich informacji może być zadaniem bardzo złożonym obliczeniowo. Proces uczenia sieci może być zatem postrzegany jako statystyczna kompresja informacji o rozkładzie do zwartej i jednocześnie bardziej przydatnej do wnioskowania bayesowskiego postaci.

Konstruowanie sieci Bayesowskiej składa się z następujących kroków:

- zdefiniowanie zmiennych,
- zdefiniowanie połączeń pomiędzy zmiennymi,
- określenie prawdopodobieństw warunkowych *a priori*
- wprowadzenie danych do sieci,
- uaktualnienie sieci,
- wyznaczenie prawdopodobieństw *a posteriori*

Uczenie sieci jest uczeniem bez nadzoru i bez wstępnej wiedzy eksperckiej, a co za tym idzie zakłada się, że na wstępie (bez znajomości próby statystycznej) wszystkie dozwolone struktury sieci są jednakowo prawdopodobne, a przy ustalonej strukturze grafu wszystkie możliwe poprawne zbiory tablic prawdopodobieństw warunkowych są jednakowo prawdopodobne.

1.1.1 Algorytm Chow-Liu

Algorytm Chow-Liu stanowi klasyczny algorytm odtwarzający kształt zależności w próbie. Algorytm nie buduje sieci Bayesowskiej, a jedynie niezororientowane drzewo zależności. Jeżeli sieć Bayesa danego rozkładu ma postać drzewa, to algorytm Chow-Liu poprawnie odtworzy jego kształt. Algorytm Chow-Liu nie sprawdza, czy rozkład zmiennych zadanej próby D spełnia powyższy warunek. Jeśli struktura zależności między zmiennymi w próbie nie jest drzewem, to algorytm znajduje najlepszą strukturę drzewiastą, która opisuje postać zależności. Należy jednak pamiętać, że w skrajnie niekorzystnych wypadkach informacja zwrócona przez algorytm może być błędna.

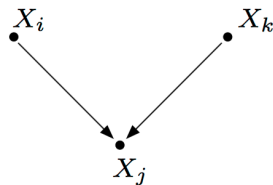
Przebieg algorytmu:

Krok 1: Niech G oznacza graf pełny o zbiorze wierzchołków tożsamym ze zbiorem atrybutów próby D . Wówczas niech T będzie nieskierowaną strukturą drzewiastą uzyskaną w wyniku zastosowania dowolnego algorytmu znajdowania minimalnego drzewa rozpinającego (MST) w grafie G z funkcją wagową określającą stopień zależności między zmiennymi.

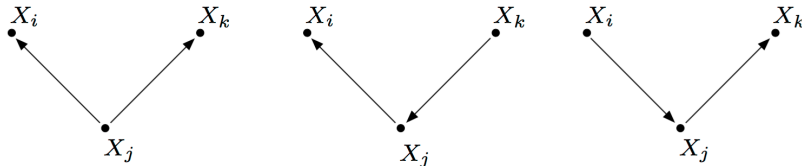
Krok 2: W drzewie T należy obrać kierunki w sposób dowolny, uzyskując wynikową strukturę sieci Bayesa BS.

W pierwszym kroku możliwe jest wykorzystanie odległości Kullback-Leiblera jako funkcji wagowej:

$$DEP(X_i, X_j) = \sum_{x_i, x_j} P(x_i, x_j) \log \frac{P(x_i, x_j)}{P(x_i)P(x_j)}$$



Rysunek 2: Dwu rodziców: $DEP(X_i, X_k) = 0 \wedge DEP(X_i, X_k|X_j) > 0$



Rysunek 3: X_j d-separuje X_i, X_k : $DEP(X_i, X_k) > 0 \wedge DEP(X_i, X_k|X_j) = 0$

Gdzie podobnie jak w warunku małe litery x_i, x_j oznaczają wartościowania zmiennych X_i, X_j a sumowanie przebiega po wszystkich wartościowaniach w próbie D .

1.2 Ukryte modele Markowa

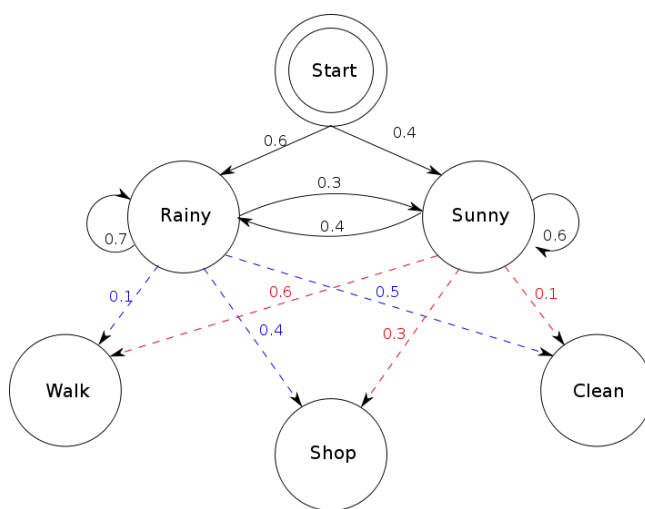
1.2.1 Wstęp

Ukryte modele Markowa (ang. *Hidden Markov Model*, w skrócie *HMM*) jest określeniem zaawansowanego modelu statystycznego znajdującego obecnie zastosowanie w wielu dziedzinach inżynierskich, szczególnie tam, gdzie analizowane są zjawiska o charakterze sekwencji losowych zdarzeń jak na przykład mowa czy gesty. Termin ten został wprowadzony i opisany matematycznie w drugiej połowie lat sześćdziesiątych ubiegłego wieku przez Bauma i Petriego, zaś swą nazwę zawdzięcza podstawie matematycznej, na której się opiera - łańcuchowi Markowa, który to określony został przez rosyjskiego matematyka Andrieja Markowa.

W rozumieniu ogólnym ukryte modele Markowa określają system zdolny z pewnym prawdopodobieństwem przewidzieć jak zachowa się modelowany obiekt w przyszłości, bazując tylko i wyłącznie na jego aktualnym stanie, gdyż nie jest przechowywana historia wyników, jakie osiągał w przeszłości. Cały system dzieli się na niewidoczne dla obserwatora ukryte stany oraz część obserwowaną (wyjście), która jest losową funkcją stanu.

Dzięki swej rozbudowanej matematycznej charakterystyce, ukryte modele Markowa sukcesywnie stosowane są jako baza teoretyczna do rozwiązywania wielu problemów, a poprawnie zastosowane w praktyce dają bardzo dobre rezultaty. Głównymi zagadnieniami, przy których są one stosowane są modele akustyczne przy rozpoznawaniu mowy, rozpoznawanie pisma ręcznego, obiektów czy gestów. Znajdują zastosowanie również szeroko w bioinformatyce, biomedycynie, w psychologii do modelowania procesów uczenia się, w finansach przy modelowaniu ryzyka na rynku obligacji, przy prognozowaniu pogody, do generowania muzyki czy do wypełniania brakujących wyrazów w zdaniach. Jako ciekawe zastosowania wymienić również można modelowanie erupcji gejzeru *Old Faithful* czy zbudowanie bota *Mark V. Shaney* podszywającego się pod zwykłego użytkownika *Usenetu* w latach osiemdziesiątych.

Jednym z głównych problemów występujących przy budowanie takiego modelu jest określenie odpowiedniego układu poprzez wyznaczenie topologii i dobranie wartości parametrów tak, aby otrzymać jak największą efektywność przy rozpoznawaniu. Klasyczne metody doboru parametrów modelu, jak algorytm *Baum-Welcha*, czy metody gradientowe, nie zapewniają znalezienia optymalnych wartości oraz wymagają poczynienia wstępnych założeń co do topologii modelu. Z tego to względu w ostatnich latach następuje znaczący wzrost zainteresowania tworzeniem nowych bądź udoskonalaniem obecnych metod budowy modelu.



Rysunek 4: Ogólny wygląd modelu Markowa.

1.2.2 Łańcuchy Markowa

Łańcuchy Markowa najprościej określić można za pomocą jednej z dostępnych definicji. Wszystkie rozważania dotyczyć będą funkcji dyskretnych w czasie.

Definicja 1. Ciąg zmiennych losowych (X_t) o wartościach w przeliczalnym zbiorze S_X (przestrzeni stanów) nazywamy łańcuchem Markowa wtedy i tylko wtedy, gdy dla każdego $t \in \mathbb{N}$ i każdego ciągu $x_1, x_1, \dots, x_t \in S_X$ mamy

$$\begin{aligned} P(X_t = x_t \mid X_{t-1} = x_{t-1}, \dots, X_2 = x_2, X_1 = x_1) = \\ = P(X_t = x_t \mid X_{t-1} = x_{t-1}) \end{aligned} \quad (1)$$

jeśli tylko

$$P(X_{t-1} = x_{t-1}, \dots, X_2 = x_2, X_1 = x_1) > 0. \quad (2)$$

Warunek w powyższej definicji określa cały charakter łańcuchów Markowa, czyli, że ewolucja procesu zależy tylko i wyłącznie od bieżącego stanu. Zatem na charakter zmiennej X_t w zadanej chwili t ma wpływ wartość procesu w chwili $t - 1$. Taki łańcuch nazywamy łańcuchem Markowa rzędu pierwszego, czyli jego stan zależy tylko od stanu poprzedniego. Istnieją także rzędy x , mówiące, że stan zależy od x -stanów poprzednich.

W wypadku, gdy prawdopodobieństwa przejścia między stanami nie zależą od momentu, w którym rozpatrywany jest proces, wtedy łańcuch określić można mianem jednorodnego w czasie. Prawdopodobieństwo takie oznacza się jako p_{ij} . Macierz kwadratowa $P = [p_{ij}]$ jednorodnego łańcucha Markowa określa się jako macierz przejść.

1.2.3 Definicja ukrytego modelu Markowa

Dla rozważań w tym rozdziale, niech:

T = długość obserwowanej sekwencji

N = liczba stanów w modelu

M = liczba obserwacji

Mając na uwadze, że ukryty model Markowa jest szczególnym przypadkiem łańcuchu Markowa, możemy teraz opisać ten układ. Przejścia między stanami opisane niech będą przy pomocy kwadratowej macierzy prawdopodobieństw $A = [a_{ij}]$, gdzie element a_{ij} jest to prawdopodobieństwo przejścia od stanu i do stanu j w następnej chwili czasowej.

$$a_{ij} = \Pr(x_t = j \mid x_{t-1} = i) \quad \text{dla} \quad 1 \leq i, j \leq N$$

Spełniona jest równość:

$$\sum_{j=1}^N a_{ij} = 1$$

Kolejną składową systemu jest macierz Π zawierająca prawdopodobieństwa wystąpienia i -tego stanu na początku sekwencji stanów.

$$\Pi_i = Pr(x_0 = i)$$

Znając obie macierze A oraz Π można obliczyć prawdopodobieństwo wygenerowania przez system sekwencji stanów $X = (x_0, x_1, \dots, x_T)$.

$$Pr(x | A, \Pi) = \Pi_{x_0} a_{x_0 x_1} a_{x_1 x_2} \dots x_{x_{T-1} x_T}$$

Tak opisana została warstwa ukryta. Model ten posiada jednakże jeszcze drugą warstwę, określaną mianami warstwy ukrytej, obserwowanej, bądź emisji. Każdemu ze stanów ukrytych przypada pewne prawdopodobieństwo b , mówiące, że w trakcie przebywania w nim, wygenerowana zostanie obserwacja O_t .

$$B = \{b_i(O_t)\}_{i=1}^N$$

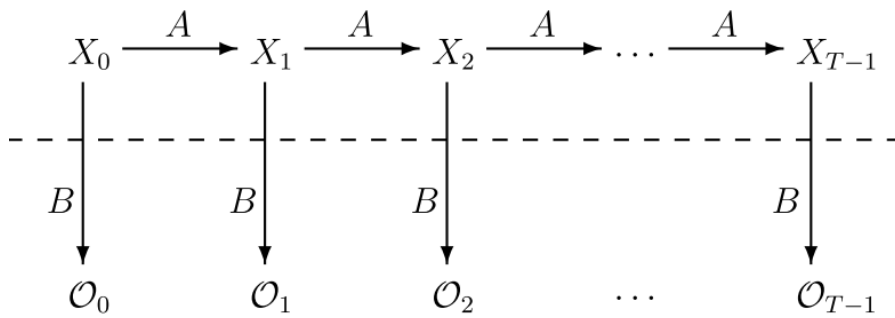
Posiadając te wszystkie informacje można określić ukryty model Markowa jako λ :

$$\lambda = (\Pi, A, B)$$

Teraz można obliczyć prawdopodobieństwo wygenerowania sekwencji obserwacji O przez system λ :

$$Pr(O | \Pi, A, B) = \sum_x P(O, x | \Pi, A, B) = \sum_x \Pi_{x_0} \prod_{t=1}^T a_{x_{t-1} x_t} b_{x_t}(O_t)$$

Ukryty model Markowa najprościej jest zobrazować przy pomocy poniższego schematu, zwanego grafem *Trellisa*, gdyż ukazuje zmienność procesu wraz z czasem. Linia kropkowaną oddzielono to co widoczne jest przez obserwatora od warstwy ukrytej. Strzałki między węzłami mówią o bezpośredniej zależności stochastycznej, zaś brak strzałki oznacza niezależność losową.



Rysunek 5: Ukryty model Markowa

Gdzie:

$\mathbf{X} = (X_0, X_1, \dots, X_{T-1})$ ukryte stany

$\mathbf{O} = (O_0, O_1, \dots, O_{T-1})$ sekwencja obserwacji

\mathbf{A} = macierz prawdopodobieństw przejść między stanami

\mathbf{B} = macierz prawdopodobieństw wystąpienia obserwacji (emisji)