Spis treści

Podstawowe definicje

Niech zmienne losowe X, Y i Z będą dyskretne.
Prawdopodobieństwo, że zmienna losowa X przyjmuje stan x i zmienna losowa Y przyjmuje stan y będziemy oznaczać poprzez: $P(X = x, Y = y)$.

- Reguła bayesa: Mamy zmienne losowe X i Y. Zachodzi następująca równość: $\underset{x \in \text{supp}(X),\, y \in \text{supp}(Y)}{\forall} P(X = x | Y = y) = \frac{P(Y=y|X=x) \cdot P(X=x)}{P(Y=y)}$.
- Graf: Grafem nazywamy parę $G = (V, E)$, gdzie V to niepusty zbiór wierzchołków (węzłów), a E to zbiór krawędzi (krawędź jest parą wierzchołków). W przypadku, gdy każda krawędź jest uporządkowaną parą wierzchołków (krawędzie posiadają swój kierunek), to mówimy o **grafie skierowanym**.

- Ścieżka: Ścieżką z węzła V_1 do V_n (oznaczamy przez $V_1 \rightarrow V_n$) nazywamy ciąg $(V_k)_{k\in\{1,2,\ldots,n\}}$, gdzie każda krawędź (V_{k-1}, V_k) należy do grafu G (dla $k \in \{2, 3, \ldots, n\}$).
- Przodkowie i potomkowie: Jeżeli istnieją dwa wierzchołki V_1 i V_n takie że: $V_1 \rightarrow V_n \wedge V_n \not\rightarrow V_1$, to wierzchołek V_1 nazywamy **przodkiem** wierzchołka V_2, natomiast wierzchołek V_2 nazywamy **potomkiem** wierzchołka V_1.
- Cykl i graf acykliczny: Cyklem nazywamy skierowaną ścieżkę, która się zaczyna i kończy na tym samym wierzchołku. Jeżeli graf nie zawiera żadnych cykli, to taki graf nazywamy grafem acyklicznym.

- Warunkowa niezależność zmiennych losowych: Mówimy, że zmienna losowa X jest niezależna od zmiennej losowej Y pod warunkiem stanu zmiennej losowej Z (oznaczamy przez: $X \perp\!\!\!\perp Y|Z$), jeżeli zachodzi następująca równość: $\forall_{x \in \text{supp}(X),\, y \in \text{supp}(Y),\, z \in \text{supp}(Z)} P(X = x, Y = y|Z = z) = P(X = x|Z = z) \cdot P(Y = y|Z = z)$ (równoważnie: $P(X = x|Y = y, Z = z) = P(X = x|Z = z)$). Jeżeli taka równość nie zachodzi, to mówimy, że zmienna losowa X jest zależna od zmiennej losowej Y pod warunkiem zmiennej losowej Z.

Mamy zmienne losowe $X_1, X_2, \ldots, X_n$. Prawdopodobieństwo, że zmienne losowe przyjmują odpowiednio stan $x_1, x_2, \ldots, x_n$ będziemy zapisywać skrótowo poprzez $P(x_1, x_2, \ldots, x_n)$.

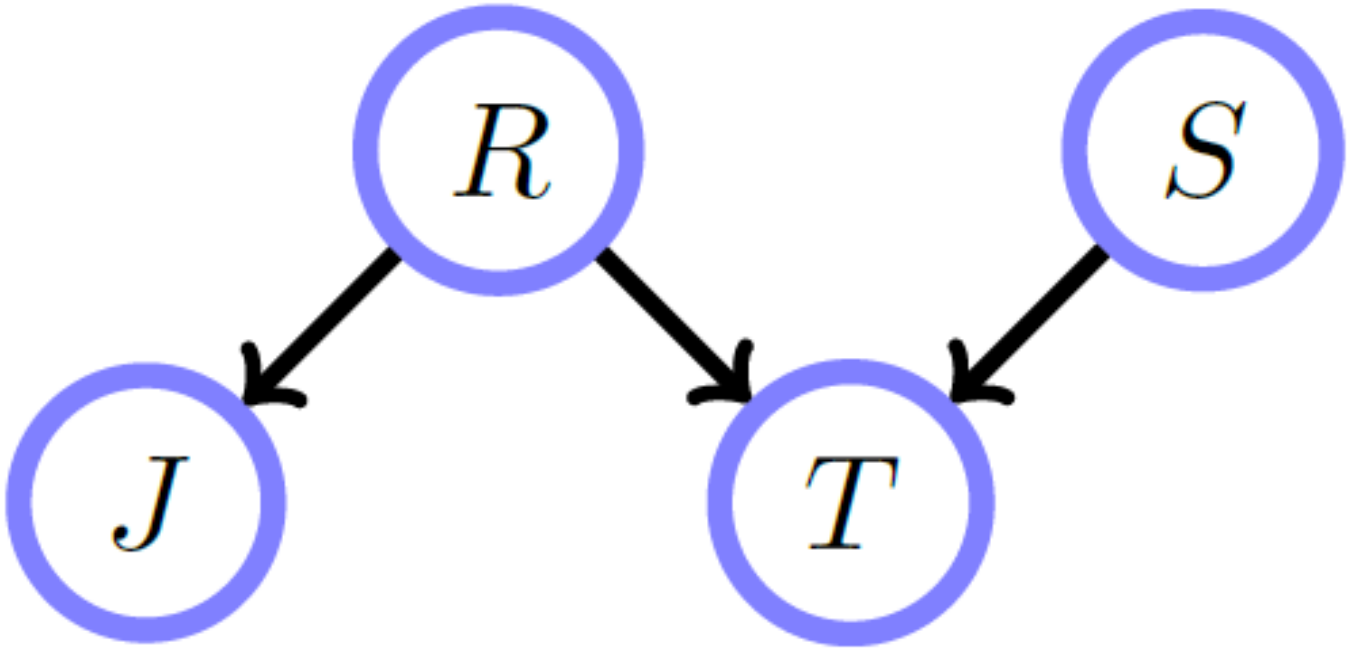
- Sieć bayesowska: Sieć bayesowska jest rozkładem o postaci: $P(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} P(x_i|R(x_i))$, gdzie $R(x_i)$ jest zbiorem rodziców zmiennej X_i. Załóżmy, że graf G jest DAG, wtedy takie prawdopodobieństwo utożsamiamy z wierzchołkiem V_i tego grafu.
- Warunek Markowa: Niech $\mathcal{P}$ będzie łącznym rozkładem zmiennych losowych zdefiniowanych na zbiorze V. Ponadto, zdefiniujmy DAG $G = (V, E)$. Mówimy, że para $(\mathcal{P}, G)$ spełnia warunek Markowa, jeżeli zachodzi: $\underset{X \in V}{\forall} X \perp\!\!\!\perp NP(X)|R(X)$, gdzie $NP(X)$, to zbiór wierzchołków, które nie są przodkami wierzchołka X, oraz $R(X)$, to zbiór wierzchołków będących rodzicami wierzchołka X.

Warto zauważyć, że jeżeli zachodzi $X \perp\!\!\!\perp NP(X)|R(X)$, to również zachodzi $\underset{P \in NP(X)}{\forall} X \perp\!\!\!\perp P|R(X)$.

Przykład prostej sieci bayesowskiej

Pewnego dnia Tracey wyszła z domu i zauważyła, że jej trawa jest mokra. Zastanawiała się czy to jest wina nocnego deszczu, czy tego, że zapomniała wyłączyć zraszaczy. Następnie zauważyła, że u jej sąsiadza Jacka, trawa też jest mokra. Taka informacja, może w pewnym stopniu wykluczyć, że jej trawa jest mokra z powodu wyłączonych zraszaczy. Zdefiniujmy binarne zmienne losowe T, J, S, R ($T = 1$, gdy trawa u Tracey jest mokra, $J = 1$, gdy u Jacka trawa jest mokra, $S = 1$, gdy zraszacze zostały włączone i $R = 1$, gdy padał deszcz w nocy). Jeżeli chcielibyśmy teraz stworzyć łączny rozkład prawdopodobieństwa, to moglibyśmy go przedstawić następująco (przykładowo):

$$P(T, J, S, R) = P(T|J, R, S) \cdot P(J, R, S) =$$
$$P(T|J, R, S) \cdot P(J|R, S) \cdot P(R, S) = P(T|J, R, S) \cdot P(J|R, S) \cdot P(R|S) \cdot P(S)$$


R
S
J
T

Możemy zauważyć, że zachodzi $P(T|J,R,S) = P(T|R,S)$ (ponieważ to czy u Tracey trawa jest mokra, zależy tylko od tego czy spadł deszcz i czy zraszacze były włączone). Możemy to teraz odnieść do warunkowej niezależności. Czyli $T \perp\!\!\!\perp J|\{R,S\}$. Ponadto, możemy zredukować pozostałe prawdopodobieństwa warunkowe:

- $P(J|R,S) = P(J|R)$ (to czy trawa u Jacka jest mokra, zależy tylko od tego, że deszcz spadł)
- $P(R|S) = P(R)$ (to czy deszcz spadł niezależy od żadnych (znanych nam) czynników)

Dzięki temu, możemy zredukować łączny rozkład prawdopodobieństwa do: $P(T,J,S,R) = P(T|R,S) \cdot P(J|R) \cdot P(R) \cdot P(S)$.

Testowanie warunkowej niezależności

Na sam początek załóżmy, że dla każdego z niżej wymienionych testów, rozważamy problem testowania:

$$H_0 : X \perp\!\!\!\perp Y | Z$$

przeciwko

$$H_1 : X \not\perp\!\!\!\perp Y | Z.$$

Test informacji wzajemnej: Jest to rodzaj testu dla zmiennych dyskretnych. Zdefiniujmy zmienne losowe X, Y i Z, gdzie $\#\,\mathrm{supp}(X) = a$ (supp to oznaczenie nośnika, a $\#$ to moc zbioru), $\#\,\mathrm{supp}(Y) = b$ oraz $\#\,\mathrm{supp}(Z) = c$. Zdefiniujmy statystykę

$MI(X, Y|Z) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c} \frac{n_{ijk}}{n} \log \frac{n_{ijk} \cdot n_{++k}}{n_{i+k} \cdot n_{+jk}}$, gdzie n to liczba obserwacji, n_{ijk} to liczba takich obserwacji, że $(X = i, Y = j, Z = k)$, natomiast oznaczenie $+$ oznacza, że idziemy po wszystkich realizacjach danej zmiennej losowej. Na przykład n_{+jk} to liczność takich obserwacji, że $(Y = j, Z = k)$. Ta statystyka, pod warunkiem prawdziwości H_0, ma asymptotyczny rozkład chi-kwadrat z $(a-1)\cdot(b-1)\cdot c$ stopniami swobody. Odrzucamy H_0 dla dużych wartości tej statystyki. Ponadto, za pomocą metody Monte - Carlo możemy wyznaczyć taką wartość t, dla której $P_{H_0}(MI(X, Y|Z) > t) = \alpha$.

Test χ^2 Pearsona: Ponownie będziemy używać tego testu dla zmiennych dyskretnych. Wszystkie oznaczenia jak wyżej. Statystyka testowa jest postaci: $\chi^2(X, Y|Z) = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{c}\frac{(n_{ijk} - m_{ijk})^2}{m_{ijk}}$, gdzie $m_{ijk} = \frac{n_{i+k} \cdot n_{+jk}}{n_{++k}}$.
Odrzucamy H_0 (na asymptotycznym poziomie istotności α), gdy $\chi^2(X, Y|Z) > q(1-\alpha, (a-1)\cdot(b-1)\cdot c)$, gdzie $q(1-\alpha, (a-1)\cdot(b-1)\cdot c)$, to kwantyl rozkładu χ^2 z $(a-1)\cdot(b-1)\cdot c$ stopniami swobody.

Algorytmy i scory w sieciach bayesowskich

Zdefiniujmy złożoność sieci: $d = \sum_{i=1}^{n}(r_i - 1) \cdot q_i$, gdzie n to liczba zmiennych losowych, r_i to liczba możliwych stanów zmiennej losowej X_i, natomiast q_i to kombinacja liczby stanów wszystkich zmiennych losowych będących rodzicami zmiennej losowej X_i. Możemy to zapisać w następujący sposób: $q_i = \prod_{X_j \in R(X_i)} r_j$. [4]

Niżej zdefiniujemy trzy najpopularniejsze (dla dyskretnych sieci bayesowskich) *scory*:

- *log-wiarogodność*: $\log \hat{L}$, gdzie $\hat{L}$ jest estymatorem funkcji wiarogodności, uzyskanym na podstawie danych.
- *AIC*: $-2 \log \hat{L} + 2d$, gdzie d, to złożoność sieci.
- *BIC*: $-2 \log \hat{L} + \log(N)d$, gdzie N to liczba obserwacji.

Niech $s(D, G)$ będzie *score'm* uzyskanym na DAG G i zbiorze danych D. W przykładach obliczeniowych będziemy korzystać z algorytmu, który dopasowuje DAG (opisujący pewną sieć bayesowską) do danych. Algorytm ten nazywa się *Hill - climbing*. Poniżej znajduje się sposób działania tego algorytmu:

1. Na sam początek rozpoczynamy od początkowego DAG G (najczęściej jest on pusty) i wyliczamy *score* $S_{\max} = S_G = s(D, G)$. Podstawiamy $G_{\max} = G$.
2. Dla każdego możliwego dodania, usunięcia lub odwrócenia krawędzi wyliczamy wartość $S_{G^*} = s(D, G^*)$. Jeżeli $S_{G^*} > S_{\max}$ i $S_{G^*} > S_G$, to podstaw $G = G^*$ i $S_G = S_{G^*}$.
3. Jeżeli $S_G > S_{\max}$, to podstaw $S_{\max} = S_G$ i $G_{\max} = G$ i wróć do punktu 2. W przeciwnym przypadku zakończ algorytm i zwróć $G_{\max}$.

Dodatkowo będziemy korzystać z metody estymacji prawdopodobieństw warunkowych, utożsamianych z wierzchołkami DAG G:

- *MLE* $\hat{P}(X_i = x_i | R(X_i) = r(X_i)) = \frac{\hat{P}(X_i=x_i, R(X_i)=r(X_i)}{\hat{P}(R(X_i)=r(X_i)} = \frac{n_{x_i,r(X_i)}}{n_{r(X_i)}}$, gdzie $n_{x_i,r(X_i)}$ to liczba obserwacji dla których zmienna losowa X_i przyjmuje wartość x_i i rodzice zmiennej losowej X_i przyjmują wektor wartości $r(X_i)$.
- *Estymator bayesowski*
  $\hat{P}(X_i = x_i | R(X_i) = r(X_i)) = \frac{\hat{P}(X_i=x_i, R(X_i)=r(X_i)}{\hat{P}(R(X_i)=r(X_i)}$, gdzie
  $\hat{P}(X_i = x_i, R(X_i) = r(X_i)) = \frac{iss}{iss+n}\pi_{X_i,R(X_i)} + \frac{n}{iss+n}\frac{n_{x_i,r(X_i)}}{N}$ oraz
  $\hat{P}(R(X_i) = r(X_i)) = \frac{iss}{iss+n}\pi_{R(X_i)} + \frac{n}{iss+n}\frac{n_{r(X_i)}}{N}$, gdzie N to liczba obserwacji, $\pi_{X_i,R(X_i)} = \frac{1}{q_i \cdot r_i}$ i $\pi_{R(X_i)} = \frac{r_i}{q_i \cdot r_i}$ (r_i i q_i jak wyżej).

Przykład obliczeniowy

Niech k oznacza liczbę obserwacji w zbiorze, którego używamy do ewaluacji. W przypadku binarnego problemu klasyfikacyjnego, zdefiniujemy miary dobroci TP (true positive), TN (true negative), FP (false positive) i FN (false negative). Niech 1 oznacza, że dany stan występuje (lub został zaklasyfikowany). Poniższa tabela definiuje wyżej wymienione miary dla i-tej obserwacji:

		Predykcja	
		1	0
Prawda	1	$tp(i)$	$fn(i)$
	0	$fp(i)$	$tn(i)$

Wtedy $TP = \sum_{i=1}^{k} tp(i)$, $FN = \sum_{i=1}^{k} fn(i)$, $TN = \sum_{i=1}^{k} tn(i)$ i $FP = \sum_{i=1}^{k} fp(i)$

Dodatkowo, zdefiniujmy miarę $TPR = \frac{TP}{TP+FN}$ i $TNR = \frac{TN}{TN+FP}$.

Mamy pewną chorobę Y. Za pomocą zbioru zmiennych objaśniających chcemy przewidzieć czy dana osoba ma tę chorobę. W skład naszego zbioru danych wchodzą geny, które są oznaczone za pomocą wszystkich kombinacji cyfr oraz liter z alfabetu angielskiego. Niech $X_{1000\times260}$ oznacza zbiór zmiennych objaśniających. Warto zauważyć, że każda zmienna, zarówno objaśniająca jak i objaśniana, jest zmienną binarną. Losowo dzielimy nasz zbiór danych, w proporcjach $50\%, 25\%, 25\%$ odpowiednio, na zbiór treningowy (na tym zbiorze będziemy budować model), zbiór walidacyjny (za pomocą tego zbioru będzie rekalibrować nasz model) i testowy (ten zbiór będzie służył do ewaluacji modeli). Miarą dobroci naszych modeli będą miary TP, TN, FP, FN, TNR, i TPR. W tym przypadku nie będziemy porównywać czasów obliczeń, ponieważ są one małe.

Rezultaty przykładu

W poniższej tabeli zamieszczono miary TP, FP, TN, FN, TPR i TNR, z podziałem na modele:

	TP	FN	TN	FP	TPR	TNR
bayes	124	36	60	30	0.775	0.667
Neural	139	21	31	59	0.869	0.344
GLM	127	33	62	28	0.794	0.689
RandF	127	33	39	51	0.794	0.433

Uzyskane wyżej wyniki mogą być wykorzystane do następującego zadania: Przeprowadziliśmy testy (oparty na modelach z wyżej) na wykrycie choroby Y. Wszystkie wyszły pozytywne. Biorąc pod uwagę fakt, że na tę chorobę jest chorych $p\%$ ludzi ($P(Y=1)$), wyliczmy prawdopodobieństwo, że jesteśmy chorzy pod warunkiem pozytywnego wyniku testu:

$$P(Y=1|M(X)=1) = \frac{P(M(X)=1|Y=1)\cdot P(Y=1)}{P(M(X)=1)} =$$

$$= \frac{P(M(X)=1|Y=1)\cdot P(Y=1)}{P(M(X)=1|Y=1)\cdot P(Y=1) + P(M(X)=1|Y=0)\cdot P(Y=0)} =$$

$$= \frac{TPR\cdot P(Y=1)}{TPR\cdot P(Y=1) + (1-TNR)\cdot P(Y=0)}$$

	$P(Y=1\|M(X)=1)$	
$P(Y=1)$	0.398	0.800
bayes	0.606	0.903
Neural	0.467	0.841
GLM	0.628	0.911
RandF	0.481	0.849

Możemy również policzyć prawdopobieństwo, że jesteśmy zdrowi, pod warunkiem negatywnego testu:

$$P(Y=0|M(X)=0) = \frac{P(M(X)=0|Y=0)\cdot P(Y=0)}{P(M(X)=0)} =$$

$$= \frac{P(M(X)=0|Y=0)\cdot P(Y=0)}{P(M(X)=0|Y=0)\cdot P(Y=0) + P(M(X)=0|Y=1)\cdot P(Y=1)} =$$

$$= \frac{TNR\cdot P(Y=0)}{TNR\cdot P(Y=0) + (1-TPR)\cdot P(Y=1)}$$

	$P(Y=0\|M(X)=0)$	
$P(Y=1)$	0.398	0.800
bayes	0.818	0.426
Neural	0.799	0.396
GLM	0.835	0.455
RandF	0.761	0.344