Statystyczna analiza danych II – zadania przygotowujące do EGZAMINU

Anna Gambin

17 stycznia 2018

Zadanie 1

Rozważmy alfabet dwuliterowy z litarami U oraz Y. Łańcuch Markowa o zbiorze stanów $\{U, Y\}$ ma następującą macierz przejścia:

$$P_{UU} = 0,9996, P_{UY} = 0,0004, P_{YU} = 0,0006, P_{YY} = 0,9994$$

Znajdź rozkład stacjonarny (π_u, π_Y) dla tego łańcucha i pokaż, że:

$$P_{UY}^{(n)} = (1 - 0,999^n)\pi_Y, P_{YU}^{(n)} = (1 - 0.999^n)\pi_U$$

Ciąg N liter początkowo miał literę U na n_u pozycjach oraz literę Y na n_y pozycjach. Ewolucja tego ciągu jest modelowana przy pomocy niezależnych kopii wyżej zdefiniowanego łańcucha Markowa, który dla każdej pozycji w ciągu startuje z odpowiedniego stanu początkowego.

Po n krokach czasowych sekwencja porównywana jest z sekwencją wyjściową i zliczana jest liczba pozycji, na których występuje niezgodność liter (oznaczmy tę liczbę przez $M(n)$). Wyprowadź wzór na $M(n)$.

Zadanie 2

Rozkład Poissona $Po(\theta)$, z parametrem $\theta > 0$, ma gęstość:

$$p_\theta(k) = \frac{e^{-\theta}\theta^k}{k!} \text{ dla } \quad k = 0, 1, 2, \ldots$$

Uzasadnij przydatną własność: dla $X \sim Po(\theta)$ mamy $E[X] = Var[X] = \theta$

Załóżmy, że niezależne zmienne losowe $X_1, X_2, \ldots X_{20}$ mają rozkład Poissona. Definiujemy estymator $\hat{\theta}_1(X_1, X_2, \ldots X_{20})$ dla parametru θ następująco:

$$\hat{\theta}_1(X_1, X_2, \ldots X_{20}) = \frac{(X_1 + X_2 + \ldots + X_{20})}{20}$$

Oblicz jego wartość oczekiwaną i wariancję, czy to jest dobry estymator (uzasadnij dlaczego)?

Rozważmy też drugi estymator zdefiniowany następująco:

$$\hat{\theta}_2(X_1, X_2, \ldots X_{20}) = (X_1 + X_2 + \ldots + X_{20})$$

Oblicz jego wartość oczekiwaną i wariancję. Czy jest on lepszy czy gorszy od pierwszego?

Zadanie 3

Pacjenci płci męskiej cierpiący na wysokie ciśnienie zostali poddani odpowiedniej terapii (dieta i ćwiczenia fizyczne) w celu jego obniżenia.

W tabeli przedstawiono wyniki badań ciśnienie krwi (w mmHg) na początku programu i po 6-ciu miesiącach terapii.

pacjent	1	2	3	4	5	6	7	8	9	10
start	141	169	158	180	147	160	175	163	148	163
po terapii	142	165	150	176	143	157	170	157	143	162

Na podstawie powyższych obserwacji przetestuj efektywność terapii w obniżaniu ciśnienia. Sformułuj odpowiednie hipotezy i zastosuj odpowiedni test statystyczny. Czy spadek ciśnienia po terapii jest statystycznie istotny (p-wartość na poziomie 1%) ?

Zadanie 4

Załóżmy, że liczba zdarzeń, które zachodzą jest modelowana zmienną losową o rozkładzie Poissona o średniej λ. Dodatkowo załóżmy, że każde zdarzenie jest (niezależnie od innych) zaobserwowane z prawdopodobieństwem p. Udowodnij, że liczba zaobserwowanych zdarzeń i odpowiednio liczba niezaobserwowanych zdarzeń może być modelowana zmienną losową o rozkładzie Poissona odpowiednio o średnich λp oraz $\lambda(1-p)$.

Zadanie 5

W ukrytym modelu Markowa o zmiennych ukrytych $X_0^N = (X_0, X_1, \ldots, X_N)$ oznaczmy stany jako $\{s_1, s_2, \ldots s_k\}$ oraz zmienne losowe modelujące emitowane symbole jako $Y_0^N = (Y_0, Y_1, \ldots, Y_N)$. Zdefiniujmy zmienną $\alpha_n(i)$ następująco:

$$\alpha_n(i) = P(Y_0^n, X_n = s_i)$$

Tak zdefiniowana zmienna spełnia następującą zależność rekurencyjną:

$$\alpha_{n+1}(i) = \sum_{j=1}^{K} \alpha_n(j) b_i(v_{n+1}) A_{ij} \tag{1}$$

- Wyjaśnij co reprezentują w powyższym wzorze A_{ij}, v_{n+1} oraz $b_i(v_{n+1})$.
- Oszacuj złożoność obliczenia wiarygodności $P(Y_0^N) = \sum_i \alpha_N(i)$ z równania (1).

Zadanie 6

Przeprowadzono niewielki eksperyment, żeby zbadać wpływ poziomu ekspresji dwóch genów (nazwijmy zmienne losowe $X1$ i $X2$) na poziom ekspresji trzeciego genu (zmienna Y). W tabeli znajdują się wyniki pomiarów ekspresji dla czterech obserwacji, przy czym ekspresja $X1$ i $X2$ nie jest mierzona dokładnie, tylko w każdym z eksperymentów gen jest albo nokautowany (wartość -1) albo jego ekspresja jest sztucznie podwyższona (wartość 1). Ekspresja Y jest mierzona dokładnie z użyciem techniki PCR.

lp.	$X1$	$X2$	Y
1	-1	-1	2.4
2	-1	1	0.4
3	1	-1	3.2
4	1	1	1.6

Zapisz model liniowy, który opisuje związek pomiędzy badanymi genami, wraz z niezbędnymi założeniami.