Egzamin SAD 2022

Dorota Celińska-Kopczyńska (DCK), Krzysztof Gogolewski (KG),
Magda Markowska, Piotr Pokarowski (PP),
Piotr Radziński (PR), Jacek Sroka (JS), Ewa Szczurek (ES)

czerwiec 2022

Zadanie 1, gr A, [Osoba do kontaktu: KG]

(2 pkt) Niech $X \in \mathbb{R}^{n \times p}$ będzie wycentrowaną macierzą danych ($\forall_j \sum_i X_{i,j} = 0$). Niech $C \in \mathbb{R}^{p \times p}$ będzie macierzą kowariancji $C = X^T X/(n-1)$.

Przypomnijmy, że jeśli $C = WLW^T$ jest diagonalizacją macierzy C, gdzie W jest macierzą wektorów własnych (każda kolumna jest wektorem własnym) i L macierzą diagonalną wartości własnych, z elementami diagonali $(L_{i,i})$ uporządkowanymi malejąco, to kolumny macierzy W nazywamy *kierunkami głównymi*, zaś j-tą kolumnę macierzy XW nazywamy j-tą *składową główną*.

Niech $X = UDV^T$ będzie rozkładem SVD macierzy X, gdzie $U \in \mathbb{R}^{n \times n}$ oraz $V \in \mathbb{R}^{p \times p}$ są macierzami ortonormalnymi, zaś $D \in \mathbb{R}^{n \times p}$ macierzą diagonalną.

Wskaż **niepoprawne** stwierdzenie spośród następujących:

- Kolumny macierzy V są *kierunkami głównymi*
- $\forall_i \, L_{i,i} = D_{i,i}^2/(n-1)$
- $rank(W) = rank(V)$

SOL Kolumny macierzy U są kolejnymi *składowymi głównymi*

Zadanie 1, gr B, [Osoba do kontaktu: KG]

(2 pkt) Niech $X \in \mathbb{R}^{n \times p}$ będzie wycentrowaną macierzą danych ($\forall_j \sum_i X_{i,j} = 0$). Niech $C \in \mathbb{R}^{p \times p}$ będzie macierzą kowariancji $C = X^T X/(n-1)$.

Przypomnijmy, że jeśli $C = WLW^T$ jest diagonalizacją macierzy C, gdzie W jest macierzą wektorów własnych (każda kolumna jest wektorem własnym) i L macierzą diagonalną wartości własnych, z elementami diagonali $(L_{i,i})$ uporządkowanymi malejąco, to kolumny macierzy W nazywamy *kierunkami głównymi*, zaś j-tą kolumnę macierzy XW nazywamy j-tą *składową główną*.

Niech $X = UDV^T$ będzie rozkładem SVD macierzy X, gdzie $U \in \mathbb{R}^{n \times n}$ oraz $V \in \mathbb{R}^{p \times p}$ są macierzami ortonormalnymi, zaś $D \in \mathbb{R}^{n \times p}$ macierza diagonalną.

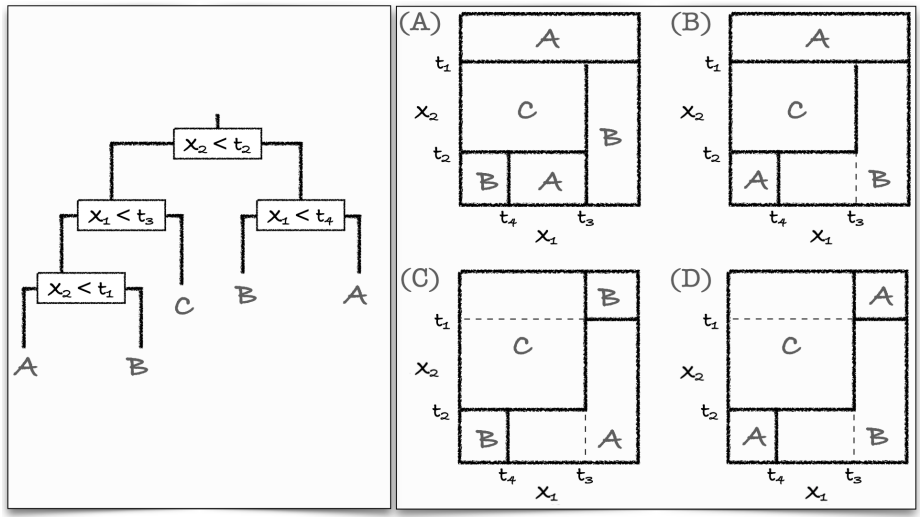
Wskaż **niepoprawne** stwierdzenie spośród następujących:

SOL Kolumny macierzy V^T są *kierunkami głównymi*

- $\forall_i \quad L_{i,i} = D^2_{i,i}/(n-1)$
- $rank(W) = rank(V)$
- Kolumny macierzy UD są kolejnymi *składowymi głównymi*

Zadanie 2 grupa A [Osoba do kontaktu: KG]

(2 pkt) Wskaż, który z czterech podziałów przestrzeni wartości parametrów X_1, X_2 odpowiada zaprezentowanemu drzewu decyzyjnemu.
Przyjmujemy, że na lewym rysunku prawe poddrzewo odpowiada za spełniony warunek zawarty w węźle; przerywane linie są tylko pomocnicze, za podział przestrzeni odpowiadają linie ciągłe; zachodzi: $t_1 > t_2$ oraz $t_3 > t_4$.

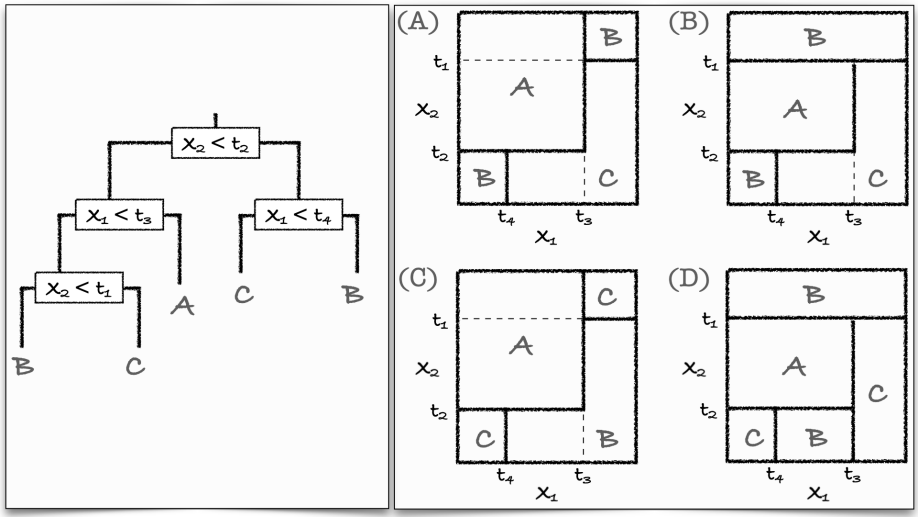


- (A)
- (B)
- (C)

SOL (D)

Zadanie 2 gr B [Osoba do kontaktu: KG]

(2 pkt) Wskaż, który z czterech podziałów przestrzeni wartości parametrów X_1, X_2 odpowiada zaprezentowanemu drzewu decyzyjnemu.
Przyjmujemy, że na lewym rysunku prawe poddrzewo odpowiada za spełniony warunek zawarty w węźle; przerywane linie są tylko pomocnicze, za podział przestrzeni odpowiadają linie ciągłe; zachodzi: $t_1 > t_2$ oraz $t_3 > t_4$.



SOL (A)

- (B)
- (C)
- (D)

Zadanie 3, gr A [Osoba do kontaktu: KG]

(2 pkt) Rozważmy model regresji liniowej z regularyzacją, w którym minimalizujemy następującą funkcję kosztu:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \frac{1}{\alpha} \|\boldsymbol{\beta}\|_2^2$$

gdzie $\boldsymbol{\beta}$ to wektor parametrów (wag) modelu, $\boldsymbol{X}$ to macierz obserwacji treningowych, $\boldsymbol{y}$ to wektor prawdziwych etykiet (wartości zmiennej objaśnianej) dla obserwacji treningowych, zaś $\alpha > 0$ to hiperparametr modelu będący liczbą rzeczywistą. Wartośc α ustalamy empirycznie w celu uniknięcia przetrenowania (*overfittingu*). Czym będzie skutkować zbyt mała wartość hiperparametru α?

- Model będzie przetrenowany

SOL Model będzie miał duże obciążenie

- Model będzie miał dużą wariancję
- Żadne z powyższych

Zadanie 3, gr B [Osoba do kontaktu: KG]

(2 pkt) Rozważmy model regresji liniowej z regularyzacją, w którym minimalizujemy następującą funkcję kosztu:

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{y}\|_2^2 + \frac{1}{\alpha} \|\boldsymbol{\beta}\|_2^2$$

gdzie $\boldsymbol{\beta}$ to wektor parametrów (wag) modelu, $\boldsymbol{X}$ to macierz obserwacji treningowych, $\boldsymbol{y}$ to wektor prawdziwych etykiet (wartości zmiennej objaśnianej) dla obserwacji treningowych, zaś $\alpha > 0$ to hiperparametr modelu będący liczbą rzeczywistą. Wartośc α ustalamy empirycznie w celu uniknięcia przetrenowania (*overfittingu*). Czym będzie skutkować zbyt duża wartość hiperparametru α?

- Model będzie miał duże obciążenie
- Model będzie niedotrenowany
- Model będzie miał małą wariancję

SOL Model będzie przetrenowany

Zadanie 4, gr A [Osoba do kontaktu: KG]

(2 pkt) Rozważmy klasyfikator binarny, który dla danej obserwacji zwraca prawdopodobieństwo p_1 posiadania etykiety 1. Model dokonuje ostatecznej predykcji etykiety wg schematu:

$$\text{Predykcja} = \begin{cases} 1 \text{ jeśli } p_1 > \tau \\ 0 \text{ jeśli } p_1 \leqslant \tau \end{cases}$$

dla danego progu τ. Model jest bardzo prosty i zawsze zwraca p_1 równe stosunkowi etykiet równych 1 do wszystkich etykiet w zbiorze treningowym (np. jeśli model był trenowany na zbiorze z etykietami $\{0, 1, 1\}$ to zawsze zwróci $p_1 = \frac{2}{3}$). τ jest hiperparametrem modelu i nie zależy od danych treningowych. Mamy następujący zbiór danych, gdzie ID to identyfikator obserwacji a y to przypisana jej etykieta:

ID	y
1	0
2	1
3	0
4	0
5	1
6	1
7	0
8	1
9	1

Model ewaluujemy w schemacie *leave-one-out cross-validation* (LOOCV) na powyższym zbiorze danych. W tym schemacie, dla jakich wartości parametru τ model będzie miał trafność (*accuracy*) równą 0?

- $\tau \in [0.375, 0.5)$
- $\tau \in (0.5, 0.625]$

SOL $\tau \in [0.5, 0.625)$

- $\tau = 0$

Zadanie 4, gr B [Osoba do kontaktu: KG]

(2 pkt) Rozważmy klasyfikator binarny, który dla danej obserwacji zwraca prawdopodobieństwo p_1 posiadania etykiety 1. Model dokonuje ostatecznej predykcji etykiety wg schematu:

$$\text{Predykcja} = \begin{cases} 1 \text{ jeśli } p_1 > \tau \\ 0 \text{ jeśli } p_1 \leqslant \tau \end{cases}$$

dla danego progu τ. Model jest bardzo prosty i zawsze zwraca p_1 równe stosunkowi etykiet równych 1 do wszystkich etykiet w zbiorze treningowym (np. jeśli model był trenowany na zbiorze z etykietami $\{0, 1, 1\}$ to zawsze zwróci $p_1 = \frac{2}{3}$). τ jest hiperparametrem modelu i nie zależy od danych treningowych. Mamy następujący zbiór danych, gdzie ID to identyfikator obserwacji a y to przypisana jej etykieta:

ID	y
1	0
2	1
3	1
4	0
5	1
6	1
7	0
8	1
9	1

Model ewaluujemy w schemacie *leave-one-out cross-validation* (LOOCV) na powyższym zbiorze danych. W tym schemacie, dla jakich wartości parametru τ model będzie miał trafność (*accuracy*) równą 0?

SOL $\tau \in [0.625, 0.75)$

- $\tau \in (0.5, 0.625)$
- $\tau \in [0.5, 0.625)$
- $\tau = 1$

Zadanie 5, gr A [Osoba do kontaktu: DCK]

(3 pkt) Dany jest model regresji liniowej:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

gdzie $\varepsilon \sim N(0, \sigma^2 I_n)$, dla którego na próbie $n = 140$ obserwacji otrzymano następujące oszacowania współczynników:

$$\hat{\beta}^T = [4, -1.5, 10, -0.5]$$

oraz następujące wartości na diagonali oszacowanej macierzy wariancji-kowariancji dla estymatora MNK $Cov(\hat{\beta})$: $[0.16, 6.25, 2.56, 0.04]$.

Które z oszacowań są statystycznie nieistotne na poziomie istotności 1%?

- $\hat{\beta}_0$ i $\hat{\beta}_3$
- $\hat{\beta}_0$, $\hat{\beta}_1$, i $\hat{\beta}_2$
- $\hat{\beta}_1$ i $\hat{\beta}_2$

SOL $\hat{\beta}_1$ i $\hat{\beta}_3$

Zadanie 5, gr B [Osoba do kontaktu: DCK]

(3 pkt) Dany jest model regresji liniowej:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon,$$

gdzie $\varepsilon \sim N(0, \sigma^2 I_n)$, dla którego na próbie $n = 140$ obserwacji otrzymano następujące oszacowania współczynników:

$$\hat{\beta}^T = [0.75, -1.2, 5.4, -3.0]$$

oraz następujące wartości na diagonali oszacowanej macierzy wariancji-kowariancji dla estymatora MNK $Cov(\hat{\beta})$ $[0.04, 1.44, 5.76, 0.09]$.

Które z oszacowań są statystycznie nieistotne na poziomie istotności 1%?

- $\hat{\beta}_0$ i $\hat{\beta}_3$

SOL $\hat{\beta}_1$ i $\hat{\beta}_2$

- $\hat{\beta}_0$, $\hat{\beta}_1$, i $\hat{\beta}_2$
- $\hat{\beta}_1$ i $\hat{\beta}_3$

Zadanie 6, gr A [Osoba do kontaktu: JS]

(3 pkt) Zadanie jest prostym przykładem ilustrującym metodę analizy głównych składowych (ang. principal component analysis, PCA). Dana jest macierz $X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$ reprezentująca 4 punkty w $\mathbb{R}^2$. Metodą PCA zredukowano wymiar tych punktów do $\mathbb{R}$. Wskaż poprawne stwierdzenie spośród następujących:

SOL empiryczna macierz kowariancji dla X to $\frac{1}{m}\begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$ dla pewnej liczby m

- w wyniku redukcji otrzymano macierz $[0, 2\sqrt{2}, -2\sqrt{2}, 0]^T$ reprezentującą 4 punkty w $\mathbb{R}$
- kierunek, wzdłuż którego dane wejściowe mają najmniejszą wariancję, wyznaczony jest przez wektor $[2, 2]$
- dane po redukcji objaśniają 90% wariancji

Zadanie 6, gr B [Osoba do kontaktu: JS]

(3 pkt) Zadanie jest prostym przykładem ilustrującym metodę analizy głównych składowych (ang. principal component analysis, PCA). Dana jest macierz $X = \begin{bmatrix} 4 & 1 \\ 2 & 3 \\ 5 & 4 \\ 1 & 0 \end{bmatrix}$ reprezentująca 4 punkty w $\mathbb{R}^2$. Metodą PCA zredukowano wymiar tych punktów do $\mathbb{R}$. Wskaż **niepoprawne** stwierdzenie spośród następujących:

- kierunek, wzdłuż którego dane wejściowe mają największą wariancję, wyznaczony jest przez wektor $[2, 2]$
- w wyniku redukcji otrzymano macierz $[0, 0, 2\sqrt{2}, -2\sqrt{2}]^T$ w $\mathbb{R}$ reprezentującą 4 punkty w $\mathbb{R}$
- wykonując redukcję straciliśmy o 20% wariancji

SOL empiryczna macierz kowariancji dla X to $\begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$

Zadanie 7, gr A [Osoba do kontaktu: DCK]

(3 pkt) Rozważmy T_1 i T_2 - dwa nieobciążone estymatory nieznanego parametru θ. Wiemy, że wariancja T_1 wynosi 0.5, a wariancja T_2 wynosi 1 oraz że estymatory są od siebie niezależne. Rozważmy teraz estymator tego samego parametru $\theta : T_3 = a * T_1 + (1 - a) * T_2$, gdzie parametr a jest liczbą z przedziału domkniętego [0,1]. Podaj wartość parametru a, która minimalizuje błąd średniokwadratowy estymatora T_3.

- 4/7

SOL 2/3

- 3/4
- 1/2

Zadanie 7, gr B [Osoba do kontaktu: DCK]

(3 pkt) Rozważmy T_1 i T_2 - dwa nieobciążone estymatory nieznanego parametru θ. Wiemy, że wariancja T_1 wynosi 1.5, a wariancja T_2 wynosi 2 oraz że estymatory są od siebie niezależne. Rozważmy teraz estymator tego samego parametru $\theta : T_3 = a * T_1 + (1 - a) * T_2$, gdzie parametr a jest liczbą z przedziału domkniętego [0,1]. Podaj wartość parametru a, która minimalizuje błąd średniokwadratowy estymatora T_3.

SOL 4/7

- 2/3
- 3/4
- 1/2

Zadanie 8, gr A [Osoba do kontaktu: ES]

(2 pkt) Rozważmy metodę lasów losowych (random forest), z parametrami: liczba drzew $B = 10$ i liczba używanych predyktorów $m = 2$. Metoda ta jest równoważna:

- Metodzie boosting dla drzew z parametrami: liczba drzew $B = 10$, parametr ściągania $\lambda = 1$ i liczba podziałów $d = 2$.
- Metodzie bagging z liczbą prób bootstrap $B = 1000$.
- Metodzie B-krotnej walidacji krzyżowej drzew decyzyjnych przycinanych do głębokości 2.

SOL Żadnej z powyższych.

Zadanie 8, gr B [Osoba do kontaktu: ES]

(2 pkt) Niech N będzie liczbą obserwacji. Rozważmy metodę k-krotnej walidacji krzyżowej. Zaznacz **nieprawdziwą** odpowiedź:

- Leave-One-Out to odmiana k-krotnej walidacji krzyżowej dla $k = N$.

SOL Walidację krzyżową możemy zastosować jedynie wtedy, gdy w zbiorze danych liczba predyktorów jest mniejsza niż liczba obserwacji N.

- Metodę k-krotnej walidacji krzyżowej możemy używać zarówno do testowania metod regresji, jak i klasyfikacji.
- Każda obserwacja zostaje wykorzystana do testowania modelu dokładnie raz, a do trenowania dokładnie k-1 razy.

Zadanie 9, grupa A [Osoba do kontaktu: ES]

(3pkt) Rozważmy metodę klastrowania hierarchicznego. Załóżmy, że:

- liczenie odległości pomiędzy parą klastrów, niezależnie od ich wielkości, zajmuje 1 s,
- ani początkowa lista odległości par klastrów, ani lista odległości nowopowstałych klastrów od pozostałych nie są zadane, należy te odległości obliczać,
- wzięcie pary klastrów o najmniejszej odległości i połączenie jej w jeden klaster zajmuje 1 s,
- w każdym kroku algorytm spędza i^2 s, gdzie i to liczba klastrów w danej iteracji i, na sortowanie listy odległości między aktualnymi klastrami.

Przy powyższych założeniach, ile sekund zajmuje wykonanie klastrowania hierarchicznego dla n elementów?

- $$\binom{n}{2} + \frac{n(n-1)}{2} + \binom{n+1}{3}$$
- $$\binom{n}{2} + \frac{n^2}{2} + \frac{n(n+1)(2n+1)}{6}$$
- $$\frac{n(n-1)}{2} - n + 1 + 2^n$$

SOL

$$\binom{n}{2} + \frac{n(n-1)}{2} + \frac{n(n+1)(2n+1)}{6} - 1$$

Zadanie 9, grupa B [Osoba do kontaktu: ES]

(3 pkt) Rozważmy metodę klastrowania hierarchicznego. Załóżmy, że:

- liczenie odległości pomiędzy parą klastrów, niezależnie od ich wielkości, zajmuje 1 s,
- ani początkowa lista odległości par klastrów, ani lista odległości nowopowstałych klastrów od pozostałych nie są zadane, należy te odległości obliczać,
- wzięcie pary klastrów o najmniejszej odległości i połączenie jej w jeden klaster zajmuje 1 s,
- w każdym kroku algorytm spędza $\binom{i}{2}$ s, gdzie i to liczba klastrów w danej iteracji i, na sortowanie listy odległości między aktualnymi klastrami.

Przy powyższych założeniach, ile sekund zajmuje wykonanie klastrowania hierarchicznego dla n elementów?

SOL

- $$\binom{n}{2} + \frac{n(n-1)}{2} + \binom{n+1}{3}$$
- $$\binom{n}{2} + \frac{n^2}{2} + \binom{n+1}{3} - 1$$
- $$\frac{n(n-1)}{2} - n + 1 + \binom{n+1}{3}$$
- $$\binom{n}{2} + \frac{n(n-1)}{2} + \frac{n(n+1)(2n+1)}{6}$$

Zadanie 10, gr A [Osoba do kontaktu: DCK]

(3 pkt) Niech X_1, X_2, X_3 będzie próbą prostą z rozkładu normalnego z nieznanymi parametrami μ i σ. Rozważmy następujące estymatory dla μ:

$$\hat{\mu}_1 = \frac{X_1+X_2+X_3}{3}, \quad \hat{\mu}_2 = \frac{2X_1+2X_2+X_3}{5}$$

Wskaż zdanie prawdziwe:

- Obydwa estymatory są obciążone.
- Wariancje estymatorów zależą od obydwu nieznanych parametrów.
- $Var(\hat{\mu}_1) > Var(\hat{\mu}_2)$ dla każdego μ i σ.

SOL $MSE(\hat{\mu}_1) \leqslant MSE(\hat{\mu}_2)$ dla każdego μ i σ.

Zadanie 10, gr B [Osoba do kontaktu: DCK]

(3 pkt) Niech X_1, X_2, X_3 będzie próbą prostą z rozkładu normalnego z nieznanymi parametrami μ i σ. Rozważmy następujące estymatory dla μ:

$$\hat{\mu}_1 = \frac{X_1+X_2-X_3}{3}, \quad \hat{\mu}_2 = \frac{2X_1+2X_2+X_3}{5}$$

Wskaż zdanie prawdziwe:

- Obydwa estymatory są obciążone.
- Wariancje estymatorów zależą od obydwu nieznanych parametrów.

SOL $Var(\hat{\mu}_1) \leqslant Var(\hat{\mu}_2)$ dla każdego μ i σ.

- $MSE(\hat{\mu}_1) > MSE(\hat{\mu}_2)$ dla każdego μ i σ.

Zadanie 11, gr A [Osoba do kontaktu: ŁR]

(3 pkt) Jaś i Małgosia postanowili oszacować wagę swoich ulubionych ciastek. Ciastka są wyrabiane ręcznie u lokalnego sprzedawcy, zdarzają się różnice w ich wielkości. Zdecydowali zatem, że każde z nich zaopatrzy się w kilka opakowań (niekoniecznie tyle samo), i na ich podstawie każde z nich obliczy symetryczny przedział ufności dla średniej wagi ciastka (modelując wagę przy użyciu rozkładu normalnego o nieznanych parametrach średniej μ i wariancji σ^2). Małgosia ma obliczyć przedział ufności na poziomie 95%, a Jaś na poziomie 99%. Jeśli założenia modelu są słuszne, to

- przedział ufności Jasia nie może być krótszy od przedziału ufności Małgosi.
- przedział ufności Jasia nie może być dłuższy od przedziału ufności Małgosi.

SOL prawdopodobieństwo zdarzenia, że otrzymane przedziały przecinają się, nie zależy od μ i σ^2.

- znając (tylko i wyłącznie) uzyskane przez Jasia i Małgosię przedziały ufności, Baba Jaga (będąca ekspertką ze statystyki matematycznej) byłaby w stanie wywnioskować, ile jest równy estymator największej wiarogodności dla σ^2 obliczony na podstawie wszystkich zakupionych przez Jasia i Małgosię ciastek.

Zadanie 11, gr B [Osoba do kontaktu: ŁR]

(3 pkt) Jaś i Małgosia postanowili oszacować wagę swoich ulubionych ciastek. Ciastka są wyrabiane ręcznie u lokalnego sprzedawcy, zdarzają się różnice w ich wielkości. Zdecydowali zatem, że każde z nich zaopatrzy się w kilka opakowań (zawierających sumarycznie tyle samo ciastek), i na ich podstawie każde z nich obliczy symetryczny przedział ufności dla średniej wagi ciastka (modelując wagę przy użyciu rozkładu normalnego o nieznanych parametrach średniej μ i wariancji σ^2). Małgosia ma obliczyć przedział ufności na poziomie 95%, a Jaś na poziomie 99%. Jeśli założenia modelu są słuszne, to

- przedział ufności Jasia nie może być krótszy od przedziału ufności Małgosi.
- przedział ufności Jasia nie może być dłuższy od przedziału ufności Małgosi.
- prawdopodobieństwo zdarzenia, że otrzymane przedziały przecinają się, zależy od co najmniej jednego z parametrów μ i σ^2

SOL znając (tylko i wyłącznie) uzyskane przez Jasia i Małgosię przedziały ufności, Baba Jaga (będąca ekspertką ze statystyki matematycznej) byłaby w stanie wywnioskować, ile jest równy estymator największej wiarogodności dla μ obliczony na podstawie wszystkich zakupionych przez Jasia i Małgosię ciastek.

Zadanie 12, gr A [Osoba do kontaktu: PP]

(3 pkt) Dany jest model liniowy $y = X\beta + \varepsilon$, gdzie $\varepsilon \sim N(0, \sigma^2 I_n)$. Załóżmy, że macierz X wymiaru $n \times p$ ma rząd p. Niech $H = (h_{ij})$ oznacza macierz daszkową oraz $\hat{y} = Hy$. Wskaż zdanie prawdziwe:

- Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}^2\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii}^2)\sigma^2$.
- Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii})^2\sigma^2$.

SOL Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii})\sigma^2$.

- Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}^2\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii})\sigma^2$.

Zadanie 12, gr B [Osoba do kontaktu: PP]

(3 pkt) Dany jest model liniowy $y = X\beta + \varepsilon$, gdzie $\varepsilon \sim N(0, \sigma^2 I_n)$. Załóżmy, że macierz X wymiaru $n \times p$ ma rząd p. Niech $H = (h_{ij})$ oznacza macierz daszkową oraz $\hat{y} = Hy$. Wskaż zdanie prawdziwe:

SOL Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii})\sigma^2$.

- Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii})^2\sigma^2$.
- Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}^2\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii}^2)\sigma^2$.
- Dla każdego i $cov(y_i, \hat{y}_i) = h_{ii}^2\sigma^2$ oraz $var(y_i - \hat{y}_i) = (1 - h_{ii})^2\sigma^2$.

Zadanie 13, gr A [Osoba do kontaktu: PP]

(3 pkt) W zadaniu dopasowania modelu liniowego otrzymano estymator najmniejszych kwadratów $\hat{\beta} = (-2.3, 1.8, 4.2, 2.1)^T$ oraz estymator wariancji $\hat{\sigma}^2 = 4$. Dodatkowo wiadomo, że $X^T X = 2I_4$. Wówczas zbiór indeksów zmiennych o najmniejszym współczynniku C_p Mallowsa jest równy

- $\{1, 2, 3, 4\}$.

- $\{3\}$.

SOL $\{1, 3, 4\}$.

- Żadna z powyższych odpowiedzi nie jest prawidłowa.

Zadanie 13, gr B [Osoba do kontaktu: PP]

(3 pkt) W zadaniu dopasowania modelu liniowego otrzymano estymator najmniejszych kwadratów $\hat{\beta} = (-2.3, 3.1, 2.1, -4.1)^T$ oraz estymator wariancji $\hat{\sigma}^2 = 9$. Dodatkowo wiadomo, że $X^T X = 2I_4$. Wówczas zbiór indeksów zmiennych o najmniejszym współczynniku C_p Mallowsa jest równy

- $\{1, 2, 3, 4\}$.
- $\{4\}$.

SOL $\{2, 4\}$.

- Żadna z powyższych odpowiedzi nie jest prawidłowa.

Zadanie 14, gr A [Osoba do kontaktu: PP]

(3 pkt) Mamy dane $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, gdzie μ nieznane, a σ znana. Rozpatrzmy dwie statystyki

$$T = \sqrt{n}\frac{\bar{X}}{S}, \quad Z = \sqrt{n}\frac{\bar{X}}{\sigma},$$

gdzie $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Oznaczmy przez $z(q)$ kwantyl standardowego rozkładu normalnego rzędu q, a przez $t(df, q)$ kwantyl rozkładu t o df stopniach swobody rzędu q. Na podstawie tych statystyk chcemy zweryfikować hipotezę $H_0 \colon \mu = 0$ vs. $H_1 \colon \mu < 0$. Wskaż prawdziwe z poniższych twierdzeń:

SOL Testy o obszarach krytycznych $W_T\{|T| > t(n-1, 1-\frac{\alpha}{2})\}$ oraz $W_Z\{|Z| > z(1-\frac{\alpha}{2})\}$ są testami na poziomie istotności α.

- Testy o obszarach krytycznych $W_T\{T < -t(n, 1-\alpha)\}$ oraz $W_Z\{Z < -z(1-\alpha)\}$ są testami na poziomie istotności α oraz test W_Z jest testem najmocniejszym na poziomie istotności α.
- Testy o obszarach krytycznych $W_T\{T < -t(n-1, 1-\alpha)\}$ oraz $W_Z\{Z < -z(1-\alpha)\}$ są testami na poziomie istotności α oraz test W_Z **nie** jest testem najmocniejszym na poziomie istotności α.
- Testy o obszarach krytycznych $W_T\{T < -t(n-1, 1-\alpha)\}$ oraz $W_Z\{Z < -z(1-\alpha)\}$ są testami na poziomie istotności α oraz test W_T ma większą moc niż test W_Z

Zadanie 14, gr B [Osoba do kontaktu: PP]

(3 pkt) Mamy dane $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, gdzie μ nieznane, a σ znana. Rozpatrzmy dwie statystyki

$$T = \sqrt{n}\frac{\bar{X}}{S}, \quad Z = \sqrt{n}\frac{\bar{X}}{\sigma},$$

gdzie $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Oznaczmy przez $z(q)$ kwantyl standardowego rozkładu normalnego rzędu q, a przez $t(df, q)$ kwantyl rozkładu t o df stopniach swobody rzędu q. Na podstawie tych statystyk chcemy zweryfikować hipotezę $H_0 \colon \mu = 0$ vs. $H_1 \colon \mu < 0$. Wskaż **nie**prawdziwe z poniższych twierdzeń:

- Testy o obszarach krytycznych $W_T\{|T| > t(n-1, 1-\frac{\alpha}{2})\}$ oraz $W_Z\{|Z| > z(1-\frac{\alpha}{2})\}$ są testami na poziomie istotności α.
- Testy o obszarach krytycznych $W_T\{T < -t(n-1, 1-\alpha)\}$ oraz $W_Z\{Z < -z(1-\alpha)\}$ są testami na poziomie istotności α oraz test W_Z jest testem najmocniejszym na poziomie istotności α.

SOL Testy o obszarach krytycznych $W_T\{T < -t(n-1, 1-\frac{\alpha}{2}))\}$ oraz $W_Z\{Z < -z(1-\frac{\alpha}{2}))\}$ są testami na poziomie istotności α oraz test W_Z jest testem najmocniejszym na poziomie istotności α.

- Testy o obszarach krytycznych $W_T\{T < -t(n-1, 1-\frac{\alpha}{2})\}$ oraz $W_Z\{Z < -z(1-\frac{\alpha}{2})\}$ są testami na poziomie istotności $\alpha/2$.

Zadanie 15, gr A [Osoba do kontaktu: PP]

(3 pkt) Niech $X_1, \ldots, X_n$ będzie próbą prostą z rozkładu o $EX_1 = 0$ i $Var(X_1) = \sigma^2$, dla $n \geqslant 4$. Rozważmy następujące estymatory wariancji.

$$S_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2, \quad S_3 = \frac{1}{n+2}\sum_{i=1}^{n}X_i^2$$

Oznaczmy przez $b(S_i)$ obciążenie estymatora S_i. Wskaż zdanie **nie**prawdziwe:

- $b(S_1) = 0$
- $b(S_2) = 0$
- $|b(S_2)| \leqslant |b(S_3)|$

SOL Jedno z powyższych zadań jest nieprawdziwe.

Zadanie 15, gr B [Osoba do kontaktu: PP]

(3 pkt) Niech $X_1, \ldots, X_n$ będzie próbą prostą z rozkładu o $EX_1 = 0$ i $Var(X_1) = \sigma^2$, dla $n \geqslant 4$, rozważmy następujące estymatory wariancji.

$$S_1 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2, \quad S_2 = \frac{1}{n}\sum_{i=1}^{n}X_i^2, \quad S_3 = \frac{1}{n+2}\sum_{i=1}^{n}X_i^2$$

Oznaczmy przez $b(S_i)$ obciążenie estymatora S_i. Wskaż zdanie prawdziwe:

- $b(S_1) \neq 0$
- $|b(S_1)| \geqslant |b(S_3)|$
- $|b(S_2)| \geqslant |b(S_3)|$

SOL $b(S_2) = 0$