Egzamin SAD 2021

Dorota Celińska-Kopczyńska (DCK), Krzysztof Gogolewski (KG), Krzysztof Koras (KK), Błażej Miasojedow (BM), Piotr Pokarowski (PP), Agnieszka Stępień-Baran (ASB), Ewa Szczurek (ES)

Czerwiec 2021

Zadanie 1 [Autor: ES, punkty: 3, gr 1, czas: 10min]

Wskaż, które zdanie jest prawdziwe

- Estymacja out-of-bag to metoda estymacji błędu testowego wykorzystująca leave-one-out validation na danych bootstrapowanych.
- Dla danego drzewa budowanego w metodzie random forest, do zbioru out-of-bag należą te cechy (predyktory), które dla tego drzewa nie zostały wylosowane ze zbioru wszystkich cech.

SOL Estymacja out-of-bag to metoda estymacji błędu testowego, którą można stosować zarówno w metodzie bagging, jak i random forest.

- Wszystkie powyższe zdania są nieprawdziwe.

Zadanie 1 [Autor: ES, punkty: 3, gr 2, czas: 10min]

Wskaż, które zdanie jest prawdziwe

SOL Estymacja out-of-bag w metodzie ranom forest wykorzystuje fakt, że dla każdej obserwacji średnio $B/3$ z budowanych B drzew nie wykorzystuje tej obserwacji do treningu.

- Out-of-bag to metoda estymacji wariancji parametrów metody bagging, wykorzystująca bootstrap danych.
- Dla każdej iteracji metody boosting, out-of-bag zawiera te cechy (predyktory) ze zbioru wszystkich cech, które nie zostały użyte do budowy drzewa w tej iteracji.
- Wszystkie powyższe zdania są nieprawdziwe.

Zadanie 2 [Autor: KK, punkty: 3, gr 1, czas: 10min]

Załóżmy, że chcemy zbudować klasyfikator przewidujący czy nasz kolega lub koleżanka będzie pływać na windsurfingu danego dnia bądź nie. Przez 8 dni obserwujemy kolegę lub koleżankę, zwracając uwagę na trzy zmienne objaśniające: siłę wiatru, nasłonecznienie oraz to, czy w okolicy zauważono rekina. Otrzymujemy dane jak w tabeli poniżej. Do klasyfikacji chcemy użyć drzewa decyzyjnego. Przyjmując, że drzewo budujemy z góry na dół, oraz używamy entropii krzyżowej jako miary "czystości" podziału danych, która zmienna objaśniająca powinna znajdować się w wierzchołku drzewa?

Wiatr	Słońce	Rekin	Było pływane
Silny	Tak	Nie	Tak
Słaby	Tak	Tak	Nie
Umiarkowany	Nie	Nie	Tak
Silny	Tak	Tak	Nie
Silny	Nie	Nie	Tak
Słaby	Tak	Nie	Tak
Umiarkowany	Tak	Nie	Nie
Umiarkowany	Nie	Nie	Nie

- Wiatr.
- Słońce.

SOL Rekin.

- Za mało informacji by stwierdzić.

Zadanie 2 [Autor: KK, punkty: 3, gr 2, czas: 10min]

Załóżmy, że chcemy zbudować klasyfikator przewidujący czy nasz kolega lub koleżanka będzie pływać na windsurfingu danego dnia bądź nie. Przez 8 dni zbieramy dane, zwracając uwagę na trzy zmienne objaśniające: siłę wiatru, nasłonecznienie oraz to, czy w okolicy zauważono rekina. Otrzymujemy dane jak w tabeli poniżej. Do klasyfikacji chcemy użyć drzewa decyzyjnego. Przyjmując, że drzewo budujemy z góry na dół, oraz używamy entropii krzyżowej jako miary "czystości" podziału danych, która zmienna objaśniająca powinna znajdować się w wierzchołku drzewa?

Wiatr	Słońce	Rekin	Było pływane
Silny	Nie	Tak	Nie
Słaby	Tak	Tak	Tak
Umiarkowany	Nie	Nie	Nie
Silny	Tak	Tak	Tak
Silny	Nie	Nie	Nie
Słaby	Nie	Nie	Tak
Umiarkowany	Tak	Nie	Tak
Umiarkowany	Nie	Nie	Nie

- Wiatr.

SOL Słońce.

- Rekin.
- Za mało informacji by stwierdzić.

Zadanie 3 [Autor: KK, punkty: 3, gr 1, czas: 10min]

Rozważmy heurystyczny klasyfikator binarny, który zawsze przewiduje dominantę etykiet (wartości zmiennej objaśnianej) tych obserwacji, na których był trenowany. Mamy podany zbiór danych jak w tabeli poniżej, gdzie ID to identyfiaktor obserwacji, a y to przypisana jej etykieta. Model ewaluujemy z użyciem 3-krotnej walidacji krzyżowej na tym zbiorze danych. Dane do podzbiorów przydzielamy po kolei, nie losowo, bez tasowania powyższego datasetu. Biorąc pod uwagę powyższe założenia, jaka będzie średnia trafność (*accuracy*) opisanego modelu?

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

- 0.56
- 0.44

SOL 0.22

- 0.33

Zadanie 3 [Autor: KK, punkty: 3, gr 2, czas: 10min]

Rozważmy heurystyczny klasyfikator binarny, który zawsze przewiduje dominantę etykiet (wartości zmiennej objaśnianej) tych obserwacji, na których był trenowany. Mamy podany zbiór danych jak w tabeli poniżej, gdzie ID to identyfikator obserwacji, a y to przypisana jej etykieta. Model ewaluujemy z użyciem 3-krotnej walidacji krzyżowej na tym zbiorze danych. Dane do podzbiorów przydzielamy po kolei, nie losowo, bez tasowania powyższego datasetu. Biorąc pod uwagę powyższe założenia, jaka będzie średnia trafność (ang. accuracy) opisanego modelu?

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

SOL 0.67

- 0.33
- 0.22
- 0.56

Zadanie 4 [Autor: DCK, punkty 3, gr 1, czas: 10min]

Dla modelu regresji logistycznej prognozującego zakwalifikowanie osoby do programu emerytalnego złota jesień"(1 – osoba jest kwalifikowalna do tego programu, 0 – nie jest) uzyskano w zależności od przyjętego prawdopodobieństwa progowego $p^\star$ różne macierze konfuzji. Korzystając jedynie z zamieszczonych macierzy konfuzji, wskaż zdanie prawdziwe:

$p^\star = 0.25$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	3291	4067
$\hat{Y} = 0$	346	1571

$p^\star = 0.5$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	1399	1026
$\hat{Y} = 0$	2238	4612

$p^\star = 0.75$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	0	0
$\hat{Y} = 0$	3637	5638

- Dla $p^\star = 0.25$ swoistość (specyficzność) wynosi 90.5% a czułość 27.9%.

SOL Spośród przedstawionych propozycji, dla $p^\star = 0.25$ osiągamy najwyższą czułość.

- Dla $p^\star = 0.75$ swoistość (specyficzność) wynosi 100% a czułości nie da się obliczyć.
- Czułość przy $p^\star = 0.5$ jest niższa niż przy $p^\star = 0.75$.

Zadanie 4 [Autor: DCK, punkty 3, gr 2, czas: 10min]

Dla modelu regresji logistycznej prognozującego zakwalifikowanie osoby do programu emerytalnego złota jesień"(1 – osoba jest kwalifikowalna do tego programu, 0 – nie jest) uzyskano w zależności od przyjętego prawdopodobieństwa progowego $p^\star$ różne macierze konfuzji. Korzystając jedynie z zamieszczonych macierzy konfuzji, wskaż zdanie prawdziwe:

$p^\star = 0.25$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	3291	4067
$\hat{Y} = 0$	346	1571

$p^\star = 0.5$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	1399	1026
$\hat{Y} = 0$	2238	4612

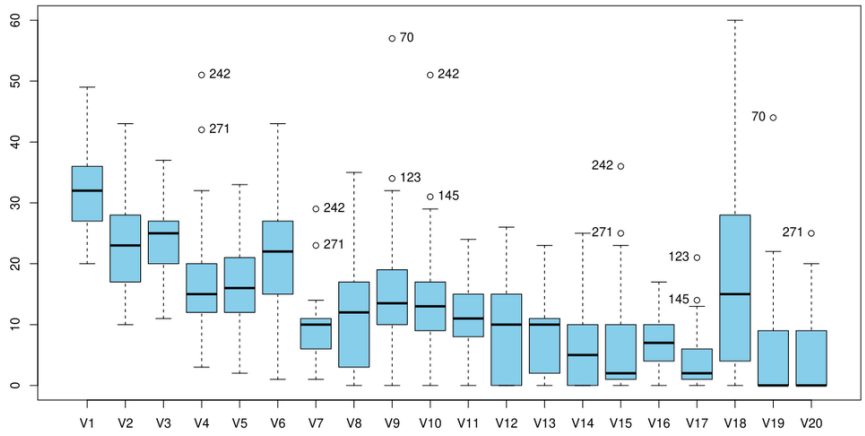
$p^\star = 0.75$	$Y = 1$	$Y = 0$
$\hat{Y} = 1$	0	0
$\hat{Y} = 0$	3637	5638

SOL Dla $p^\star = 0.25$ czułość wynosi 90.5% a swoistość (specyficzność) 27.9%.

- Spośród przedstawionych propozycji, dla $p^\star = 0.25$ osiągamy najniższą czułość.
- Dla $p^\star = 0.75$ swoistość (specyficzność) wynosi 100%, a czułości nie da się obliczyć.
- Czułość przy $p^\star = 0.5$ jest niższa niż przy $p^\star = 0.75$.

Zadanie 5 [Autor: DCK, punkty 3, gr 1, czas: 10min]

Na rysunku przedstawiono wykresy pudełkowe zmiennych wybranych do analizy PCA z pewnego zbioru danych. Jakie wnioski istotne dla analizy PCA wynikają z analizy rysunku?



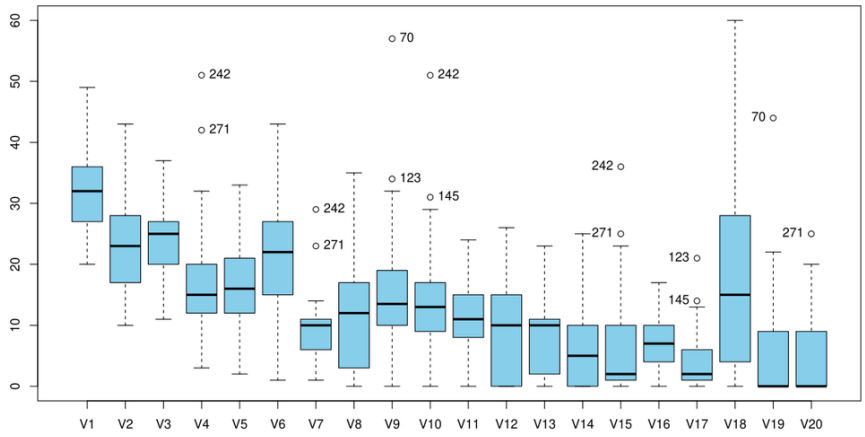
- W oparciu o rysunek nie dostrzegamy obecności obserwacji odstających
- Dostrzegamy kilka podejrzanych obserwacji, ale bez zbadania dźwigni i standaryzowanych reszt nie jesteśmy w stanie ocenić, czy powinny być usunięte

SOL Dostrzegamy obserwacje odstające, których obecność mogłaby wpłynąć na wyniki analizy PCA

- Rysunek jest nieprzydatny, bo zmienne nie zostały przeskalowane. Dopiero po przeskalowaniu zmiennych moglibyśmy zbadać, czy istnieją obserwacje odstające

Zadanie 5 [Autor: DCK, punkty 3, gr 2, czas: 10min]

Na rysunku przedstawiono wykresy pudełkowe zmiennych wybranych do analizy PCA z pewnego zbioru danych. Jakie zalecenia dla PCA wynikają z analizy rysunku?



- W oparciu o rysunek nie jesteśmy w stanie stwierdzić, czy należy skalować zmienne, potrzebujemy statystyk opisowych próby.
- Skalowanie usuwa efekt różnych jednostek, a w naszym wypadku wszystkie zmienne są mierzone na tej samej skali – nie ma więc potrzeby skalowania zmiennych.

SOL Wariancje zmiennych nie są takie same, należy przeskalować zmienne.

- W zbiorze jest za mało zmiennych, aby zastosować PCA.

Zadanie 6 [Autor: DCK, punkty 3, gr 1, czas: 10min]

Na zlecenie pewnej sieci sklepów przeprowadzono badanie ankietowe chęci zakupu nowego wariantu lodów czekoladowych znanej marki. Na podstawie zebranych danych oszacowano prosty model regresji logistycznej (1 – chce kupić, 0 – nie chce kupić), jako zmienne niezależne wykorzystując płeć respondenta (1 – mężczyzna, 0 – kobieta), wiek respondenta, liczbę dzieci w gospodarstwie domowym respondenta. Poniżej fragment wydruku z programu R.

	Estimate	Std. Error	Z value	Pr(> |z|)
(Intercept)	5.137627	0.594998	8.635	< 2e-16
plec	-2.756819	0.212026	-13.002	< 2e-16
wiek	-0.037267	0.008195	-4.547	5.43e-06
Liczba_dzieci	0.001528	0.002353	0.649	0.5160

Zgodnie z tym modelem:

Wskazówka: Obliczenia wykonaj z dokładnością do czwartego miejsca po przecinku.

- Zwiększenie liczby dzieci w gospodarstwie domowym o jedno (przy pozostałych wartościach niezmienionych) zwiększa prawdopodobieństwo chęci zakupu badanego wariantu lodów o 0.0015, efekt ten jest jednak nieistotny statystycznie
- Wśród bezdzietnych mężczyzn, 41-letni mają o około 7.2% większe prawdopodobieństwo chęci zakupu badanego wariantu lodów niż 31-letni

SOL Wśród osób mających 25 lat i jedno dziecko, mężczyzn cechuje o około 17.5 punkta procentowego niższe prawdopodobieństwo chęci zakupu badanego wariantu lodów w porównaniu do kobiet

- Prawdopodobieństwo chęci zakupu badanego wariantu lodów dla kobiety 40-letniej mającej dwójkę dzieci wynosi około 0.87

Zadanie 6 [Autor: DCK, punkty 3, gr 2, czas: 10min]

Na zlecenie pewnej sieci sklepów przeprowadzono badanie ankietowe chęci zakupu nowego wariantu lodów czekoladowych znanej marki. Na podstawie zebranych danych oszacowano prosty model regresji logistycznej (1 – chce kupić, 0 – nie chce kupić), jako zmienne niezależne wykorzystując płeć respondenta (1 – mężczyzna, 0 – kobieta), wiek respondenta, liczbę dzieci w gospodarstwie domowym respondenta. Poniżej fragment wydruku z programu R.

	Estimate	Std. Error	Z value	Pr(> |z|)
(Intercept)	5.137627	0.594998	8.635	< 2e-16
plec	-2.756819	0.212026	-13.002	< 2e-16
wiek	-0.037267	0.008195	-4.547	5.43e-06
Liczba_dzieci	0.001528	0.002353	0.649	0.5160

Zgodnie z tym modelem:

Wskazówka: Obliczenia wykonaj z dokładnością do czwartego miejsca po przecinku.

- Zwiększenie liczby dzieci w gospodarstwie domowym o jedno (przy pozostałych wartościach niezmienionych) zwiększa prawdopodobieństwo chęci zakupu badanego wariantu lodów o 0.0015, efekt ten jest jednak nieistotny statystycznie

SOL Bezdzietni 31-letni i 41-letni mężczyźni różnią się prawdopodobieństwem chęci zakupu badanego wariantu lodów o około 7.2 punkta procentowego

- Wśród osób mających 25 lat i jedno dziecko, mężczyzn cechuje o około 17.5 punkta procentowego wyższe prawdopodobieństwo chęci zakupu badanego wariantu lodów.
- Prawdopodobieństwo chęci zakupu lodów dla mężczyzny 43-letniego mającego trójkę dzieci wynosi około 0.63

Zadanie 7 [Autor: PP, punkty: 3, gr 1, czas: 10min]

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 $X^T X = 2I_4$. Wskaż zdanie prawdziwe:

- Zbiór zmiennych o najmniejszym AIC jest równy $\{1, 2, 3, 4\}$.
- Zbiór zmiennych o najmniejszym AIC jest równy $\{3\}$.

SOL Zbiór zmiennych o najmniejszym AIC jest równy $\{1, 3, 4\}$.

- Wszystkie powyższe zdania są nieprawdziwe.

Zadanie 7 [Autor: PP, punkty: 3, gr 2, czas: 10min]

W zadaniu dopasowania modelu liniowego otrzymano estymator najmniejszych kwadratów $\hat{\beta} = (-2.3, 3, 2.1, -4.1)'$ oraz estymator wariancji $\hat{\sigma}^2 = 4$. Dodatkowo wiadomo, że $X'X = I$. Wskaż zdanie prawdziwe:

- Zbiór zmiennych o najmniejszym C_p Mallowsa jest równy $\{1, 2, 3, 4\}$.

SOL Zbiór zmiennych o najmniejszym C_p Mallowsa jest równy $\{2, 4\}$.

- Zbiór zmiennych o najmniejszym C_p Mallowsa jest równy $\{4\}$.
- Wszystkie powyższe zdania są nieprawdziwe.

Zadanie 8 [Autor: PP, punkty: 3, gr 1, czas: 10min]

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

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

- Dla każdego $i = 1, \ldots, n$ $cov(y_i, \hat{y}_i) = h_{ii}\sigma^2$ oraz $var(\hat{y}_i) = h_{ii}^2\sigma^2$.
- Wszystkie powyższe zdania są nieprawdziwe.

Zadanie 8 [Autor: PP, punkty: 3, gr 2, czas: 10min]

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 **nieprawdziwe**:

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

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

- Dokładnie jedno z powyższych zdań jest nieprawdziwe.

Zadanie 9 [Autor: KG, punkty: 3, gr 1, czas:10min]

Załóżmy, że mamy zadaną macierz danych D o d kolumnach odpowiadających cechom obserwacji. Załóżmy też, że żadne dwa wektory własne w macierzy kowariancji dla macierzy D nie mają tej samej wartości własnej. Która z poniższych odpowiedzi jest **nieprawidłowa** w przypadku analizy głównych składowych (PCA)?

- Dodanie do każdej obserwacji w danych cechy, której wartość stale wynosi 1 nie zmienia wyników wykonywania PCA, poza tym, że użyteczne wektory składowych głównych mają na końcu dodatkowe 0 i jest jedna dodatkowa składowa o wartości własnej zero.
- Użycie PCA do rzutowania punktów d-wymiarowych na j głównych współrzędnych, a następnie ponowne użycie PCA do rzutowania tych j-wymiarowych współrzędnych na k głównych współrzędnych ($d > j > k$), zawsze prowadzi do tego samego wyniku (dokładnie tak jakby użyć PCA do rzutowania punktów d-wymiarowych bezpośrednio na k głównych współrzędnych).
- Jeśli wykonasz dowolny obrót danych (jako grupy punktów d-wymiarowej przestrzeni cech), największa wartość własna macierzy kowariancji danych nie zmieni się

SOL Jeśli wykonasz dowolny obrót danych (jako grupy punktów d-wymiarowej przestrzeni cech), kierunki komponentów głównych nie zmienią się.

Zadanie 9 [Autor: KG, punkty: 3, gr 2, czas: 10min]

Załóżmy, że mamy zadaną macierz danych D o d kolumnach odpowiadających cechom obserwacji. Załóżmy też, że żadne dwa wektory własne w macierzy kowariancji dla macierzy D nie mają tej samej wartości własnej. Która z poniższych odpowiedzi jest **nieprawidłowa** w przypadku analizy głównych składowych (PCA)?

- Dodanie do każdej obserwacji w danych cechy, której wartość stale wynosi 1 nie zmienia wyników wykonywania PCA, poza tym, że dotychczasowe wektory składowych głównych mają na końcu dodatkowe 0 i jest jedna dodatkowa składowa o wartości własnej zero.

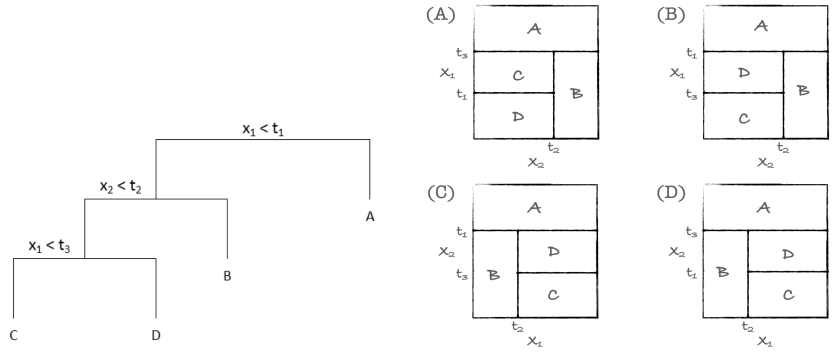
- Użycie PCA do rzutowania punktów d-wymiarowych na j głównych współrzędnych, a następnie ponowne użycie PCA do rzutowania tych j-wymiarowych współrzędnych na k głównych współrzędnych ($d > j > k$), zawsze prowadzi do tego samego wyniku (dokładnie tak jakby użyć PCA do rzutowania punktów d-wymiarowych bezpośrednio na k głównych współrzędnych).

SOL Jeśli wykonasz dowolny obrót danych (jako grupy punktów d-wymiarowej przestrzeni cech), największa wartość własna macierzy kowariancji danych zmieni się.

- Jeśli wykonasz dowolny obrót danych (jako grupy punktów d-wymiarowej przestrzeni cech), kierunki komponentów głównych zmienią się.

Zadanie 10 [Autor: KG, punkty: 2, gr 1, czas:10min]

Wskaż, który z czterech podziałów przestrzeni wartości parametrów X_1, X_2 odpowiada zaprezentowanemu drzewu decyzyjnemu.



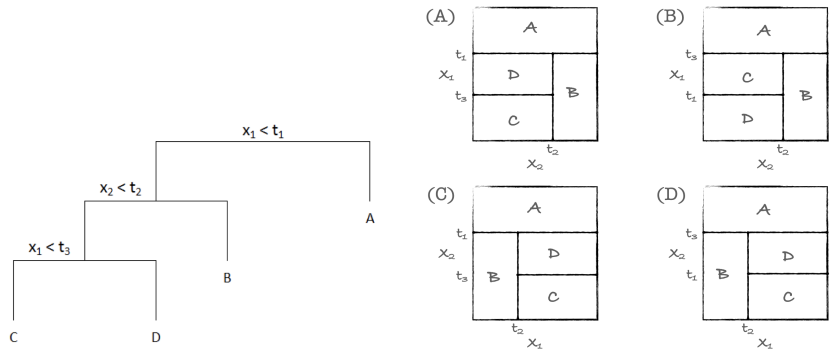
- A

SOL B

- C
- D

Zadanie 10 [Autor: KG, punkty: 2, gr 2, czas:10min]

Wskaż, który z czterech podziałów przestrzeni wartości parametrów X_1, X_2 odpowiada zaprezentowanemu drzewu decyzyjnemu.



SOL A

- B

- C
- D

Zadanie 11 [Autor: ES, punkty: 3, gr 1, czas:10min]

Mamy dane D, w których mamy n obserwacji oraz p predyktorów, a także metodę random forest i jej hiperparametry: m (liczba predyktorów losowanych dla każdego drzewa) oraz T (liczba budowanych drzew). Rozważmy kilka modeli random forest, wytrenowanych na danych D dla następujących wartości hiperparametrów: $m = p$, $T = 100$ (model oznaczony $M_{p,100}$), $m = \sqrt{p}$, $T = 100$ (model oznaczony $M_{\sqrt{p},100}$), $m = \frac{p}{2}$, $T = 100$ (model oznaczony $M_{\frac{p}{2},100}$), a także $m = p$, $T = 1$ (model oznaczony $M_{p,1}$). Wskaż, które zdanie jest **nieprawdziwe**

- $M_{p,100}$ odpowiada metodzie bagging, w której konstruuje się 100 drzew na danych bootstrapowanych z D.
- $M_{p,1}$ to drzewo decyzyjne wytrenowane na danych bootstrapowanych z D.
- Nie ma pewności, że $M_{\sqrt{p},100}$ będzie miał mniejszy błąd testowy niż $M_{\frac{p}{2},100}$.

SOL Co najmniej jedno z powyższych zdań jest nieprawdziwe.

Zadanie 11 [Autor: ES, punkty: 3, gr 2, czas:10min]

Mamy dane D, w których mamy n obserwacji oraz p predyktorów, a także metodę random forest i jej hiperparametry: m (liczba predyktorów losowanych dla każdego drzewa) oraz T (liczba budowanych drzew). Rozważmy kilka modeli random forest, wytrenowanych na danych D dla następujących wartości hiperparametrów: $m = p$, $T = 100$ (model oznaczony $M_{p,100}$), $m = \sqrt{p}$, $T = 100$ (model oznaczony $M_{\sqrt{p},100}$), $m = \sqrt{p}$, $T = 10000$ (model oznaczony $M_{\sqrt{p},10000}$), a także $m = p$, $T = 1$ (model oznaczony $M_{p,1}$). Wskaż, które zdanie jest prawdziwe

- Dla każdego D, model $M_{\sqrt{p},10000}$ będzie miał mniejszy błąd treningowy niż $M_{\sqrt{p},100}$.
- Dla każdego D, model $M_{\sqrt{p},10000}$ będzie miał większy błąd testowy niż $M_{\sqrt{p},100}$.

SOL $M_{p,1}$ odpowiada metodzie bagging, w której konstruuje się 1 drzewo na danych bootstrapowanych z D.

- $M_{p,1}$ to drzewo decyzyjne wytrenowane na danych D.

Zadanie 12 [Autor: BM, punkty: 3, gr 1, czas:10min]

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 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 12 [Autor: BM, punkty: 3, gr 2, czas:10min]

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 13 [Autor: BM, punkty: 3, gr 1, czas:10min]

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

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

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

SOL $b(S_1) = 0$

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

Zadanie 13 [Autor: BM, punkty: 3, gr 1, czas: 10min]

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

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

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

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

SOL $b(S_2) = 0$

Zadanie 14 [Autor: ASB, punkty: 3, Lab 6 i 7, czas: 10min]

Niech X_1, X_2, X_3 będzie próbą prostą z rozkładu Poissona z nieznanym parametrem $\lambda > 0$. Rozważmy następujące estymatory dla λ:

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

Wskaż zdanie prawdziwe:

- Obydwa estymatory są obciążone.
- Wariancje estymatorów nie zależą od parametru λ.
- $Var(\hat{\lambda}_1) > Var(\hat{\lambda}_2)$

SOL $MSE(\hat{\lambda}_1) < MSE(\hat{\lambda}_2)$ dla każdego λ

Zadanie 14 [**Autor: ASB, punkty: 3, Lab 6 i 7, czas: 10min**]

Niech X_1, X_2, X_3 będzie próbą prostą z rozkładu Poissona z nieznanym parametrem $\lambda > 0$. Rozważmy następujące estymatory dla λ:

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

Wskaż zdanie prawdziwe:

- Obydwa estymatory są obciążone.
- Wariancje estymatorów nie zależą od parametru λ.

SOL $Var(\hat{\lambda}_1) < Var(\hat{\lambda}_2)$

- $MSE(\hat{\lambda}_1) < MSE(\hat{\lambda}_2)$ dla każdego λ

Zadanie 15 [**Autor: ASB, punkty: 3, Lab 6 i 7, czas: 10min**]

Wysunięto hipotezę, że stopień wyprania tkaniny wełnianej płatkami mydlanymi jest wyższy od stopnia wyprania sulfapolem. W celu sprawdzenia tej hipotezy wykonano pomiary stopnia wyprania 10 wycinków tkaniny pranej płatkami, otrzymując w procentach wyniki: 74.8, 75.1, 73.0, 72.8, 76.2, 74.6, 76.0, 73.4, 72.9, 71.6 oraz 7 wyników prania sulfapolem, otrzymując: 56.9, 57.8, 54.6, 59.0, 57.1, 58.2, 57.6. Zakładając, że stopień wyprania tkaniny ma rozkład normalny chcemy zweryfikować wysuniętą hipotezę na poziomie istotności $\alpha = 0.05$. Wskaż zdanie prawdziwe:

- Próby nie są równoliczne, więc nie można sięgnąć po test t.
- Hipoteza alternatywna jest hipotezą dwustronną.
- Wartość empiryczna statystyki testowej nie należy do zbioru odrzucenia.

SOL P-wartość jest mniejsza od przyjętego poziomu istotności.

Zadanie 15 [**Autor: ASB, punkty: 3, Lab 6 i 7, czas: 10min**]

Wysunięto hipotezę, że stopień wyprania tkaniny wełnianej płatkami mydlanymi jest istotnie różny od stopnia wyprania sulfapolem. W celu sprawdzenia tej hipotezy wykonano pomiary stopnia wyprania 10 wycinków tkaniny pranej płatkami, otrzymując w procentach wyniki: 74.8, 75.1, 73.0, 72.8, 76.2, 74.6, 76.0, 73.4, 72.9, 71.6 oraz 7 wyników prania sulfapolem, otrzymując: 56.9, 57.8, 54.6, 59.0, 57.1, 58.2, 57.6. Chcemy zweryfikować wysuniętą hipotezę na poziomie istotności $\alpha = 0.05$. Wskaż zdanie prawdziwe:

- Próby nie są równoliczne, więc nie można sięgnąć po test t.
- Hipoteza alternatywna jest hipotezą jednostronną.

SOL Wartość empiryczna statystyki testowej należy do zbioru odrzucenia.

- P-wartość jest większa od przyjętego poziomu istotności.