

1. Oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio.

Pytanie 1

Poprawnie

Punkty: 1,00 z 1,00

Mamy zbiór kulek niebieskich i czerwonych. Są 3 niebieskie i 7 czerwonych. Ile wynosi entropia zbioru? Wynik podaj z dokładnością do dwóch miejsc po przecinku.

Odpowiedź:

0,88



Ile wynosi entropia jeśli mamy zbiór 10 niebieskich kulek i ani jednej czerwonej kulki?

Odpowiedź:

0



Pytanie 17

Zakończone

Punkty maks.: 5,00

Dla zadanych danych oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio dla atrybutu **app popularity**.

id	app popularity	features	usability	decision
1	high	5	low	no
2	low	7	high	yes
3	medium	7	medium	yes
4	high	3	low	no
5	low	7	high	yes
6	medium	3	medium	no
7	low	7	high	yes
8	medium	5	medium	yes
9	high	7	low	no
10	medium	5	high	yes

$$\text{entropia} = -4/10 \cdot \log_2 4/10 - 6/10 \cdot \log_2 6/10 = 0,966 \text{ (te wartości są z atrybutu decyzyjnego)}$$

entropia warunkowa=

$$\text{high} = -0/3 \cdot \log_2 0/3 - 3/3 \cdot \log_2 3/3 = 0$$

$$\text{low} = -0/3 \cdot \log_2 0/3 - 3/3 \cdot \log_2 3/3 = 0$$

$$\text{med} = -1/4 \cdot \log_2 1/4 - 3/4 \cdot \log_2 3/4 = 0,81$$

$$\text{entropia warunkowa} = 0^3/10 + 0^3/10 + 0,81^4/10 = 0,324 \text{ (te wartości są z high*ilocz high/10 + low*ilocz low/10 + med*ilocz med/10)}$$

$$\text{information gain} = \text{entropia} - \text{entropia warunkowa} = 0,642$$

$$\text{intrinsic information} = -3/10 \cdot \log_2 3/10 - 3/10 \cdot \log_2 3/10 - 4/10 \cdot \log_2 4/10 = 1,566 \text{ (te wartości to -ilocz wystąpień high/10 * log2 ilosc wystąpień high/10 - ilocz wystąpień low/10 * log2 ilosc wystąpień low/10 - ilocz wystąpień med/10 * log2 ilosc wystąpień med/10)}$$

$$\text{gain ratio} = \text{information gain} / \text{intrinsic information} = 0,642 / 1,566 = 0,41$$

Dla zadanych danych oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio dla atrybutu **hotel rating**.

id	hotel rating	rooms	location	decision
1	high	50	downtown	yes
2	medium	30	suburb	no
3	low	10	downtown	yes
4	high	10	countryside	no
5	medium	50	downtown	yes
6	low	10	suburb	no
7	high	30	downtown	yes
8	low	30	countryside	yes
9	medium	50	suburb	no
10	high	50	countryside	yes

Tekst odpowiedzi Pytanie 17

$$H(S) = -(6/10 \log_2 (6/10) + 4/10 \log_2 (4/10)) = 0,971$$

$$H(S|H=h) = -(3/10 \log_2 3/10 + 1/10 \log_2 1/10 + 2/10 \log_2 2/10) = 1,3177$$

$$IG(S) = H(S) - H(S|H) = 0,971 - 1,3177 = -0,3467$$

$$II = -(3/4 \log_2 3/4 + 1/3 \log_2 1/3 + 2/3 \log_2 2/3) = 1,0566$$

$$GR = IG/II = -0,3467 / 1,0566 = 0,3281$$

Dla zadanych danych oblicz entropię, entropię warunkową, information gain, intrinsic information, gain ratio dla atrybutu **location**.

id	hotel rating	rooms	location	decision
1	high	50	downtown	yes
2	medium	30	suburb	no
3	low	10	downtown	yes
4	high	10	countryside	no
5	medium	50	downtown	yes
6	low	10	suburb	no
7	high	30	downtown	yes
8	low	30	countryside	yes
9	medium	50	suburb	no
10	high	50	countryside	yes

Tekst odpowiedzi Pytanie 17

$$\text{entropia}(S) = -(6/10)\log_2(6/10)-(4/10)\log_2(4/10) = 0.97$$

downtown: y, y, y, y

suburb: n, n, n

countryside: n, y, y

$$\text{entropia}(\text{location}=\text{downtown}) = 0$$

$$\text{entropia}(\text{location}=\text{suburb}) = 0$$

$$\text{entropia}(\text{location}=\text{countryside}) = -(1/3)\log_2(1/3)-(2/3)\log_2(2/3) = 0.92$$

$$\text{entropia}(S|\text{location}) = (4/10)*0+(3/10)*0+(3/10)*0.92 = 0.28$$

$$\text{IG}(S,\text{location}) = \text{entropia}(S)-\text{entropia}(S|\text{location}) = 0.97-0.28 = 0.69$$

$$\text{II}(S,\text{location}) = -(4/10)\log_2(4/10)-(3/10)\log_2(3/10)-(3/10)\log_2(3/10) = 1.57$$

$$\text{GR}(S,\text{location}) = \text{IG}(S,\text{location})/\text{II}(S,\text{location}) = 0.69/1.57 = 0.44$$

Pytanie 12

Poprawnie

Punkty: 1,00 z 1,00

Przyrost informacji (ang. info gain) jest zdefiniowany jako (A --- zbiór przykładów uczących, s --- atrybut użyty do podziału w węźle drzewa; Ent symbol obliczenia entropii informacji):

Wybierz jedną odpowiedź:

- a. $\text{Gain}(A,s)=\text{Ent}(s)-\text{Ent}(A|s)$;
- b. $\text{Gain}(A,s)=\text{Ent}(A)-\text{Ent}(A|s)$ ✓
- c. $\text{Gain}(A,s)=\text{Ent}(A|s)-\text{Ent}(A)$
- d. $\text{Gain}(A,s)=\text{Ent}(s)-\text{Ent}(s|A)$

Poprawna odpowiedź to: $\text{Gain}(A,s)=\text{Ent}(A)-\text{Ent}(A|s)$

Pytanie 15

Poprawnie

Punkty: 1,00 z 1,00

Miara entropii informacji stosowana w indukcyjnym uczeniu się z przykładów w przypadku klasyfikacji binarnej przyjmuje wartości (założ podstawę 2 w logarytmie):

Wybierz jedną odpowiedź:

- a. żadna z odpowiedzi nie jest prawdziwa
- b. z przedziału [0, 1], ✓
- c. dowolne
- d. większa od 1,
- e. z przedziału [0,2*k], gdzie k jest liczbą klas

2. Na podstawie powyższych wartości określ, który atrybut będzie najlepszy do wykonania podziału.

BRAK

3. Potencjalne problemy z information gain i gain ratio.

BRAK

4. W jaki sposób należy radzić sobie z wartościami ciągłymi na atrybutach?

BRAK

5. W jaki sposób radzić sobie z brakującymi wartościami w zbiorze danych?

Pytanie 2 | Zakończone Punkty maks.: 3,00 [Oflaguj pytanie](#)

Wymień i wyjaśnij 5 sposobów radzenia sobie z brakującymi danymi w zbiorze.

Tekst odpowiedzi Pytanie 2

1. Ignorowanie przykładów z brakami - rozwiązuje problem ale można stracić dużo danych
2. Wypełnianie braków najczęściej występującymi wartościami atrybutów (globalnie) - dobre, ale wartość może być zbyt ogólna
3. Wypełnianie braków najczęściej występującymi wartościami atrybutów w obrebie danej klasy - lepiej zauważa niż branie globalnej wartości
4. Wypełnianie braków najczęściej występującymi wartościami atrybutów i oznaczanie ich jako missing
5. Uznawanie wartości "null" jako odseparowana wartość

6. Źródła niespójności w danych.

Pytanie **2**

Zakończone

Punkty maks.: 3,00

Opisz czym jest **niespójność** w danych. Jakie są dwa główne typy szumu w danych?

niespójność w danych to sytuacja, gdy mamy np. 2 identyczne obserwacje/obiekty/wiersze ale zostały one przydzielone do różnych klas decyzyjnych. Może to być spowodowane np. brakującymi atrybutami albo złą klasyfikacją z powodu "bliskich" granic pomiędzy klasami głównymi typami szumów są: szum etykiet i szum atrybutów

7. Przyczyny overfittingu.

Pytanie **4**

Poprawnie

Punkty: 1,00 z 1,00

Czym jest **overfitting**?

- a. To sytuacja, w której model jest niedostosowany do danych, przez co osiąga niską skuteczność zarówno na zbiorze treningowym, jak i testowym.
- b. To zjawisko, w którym model jest zbyt dobrze dopasowany do danych treningowych, co prowadzi do niskiej skuteczności na nowych danych. ✓
- c. To metoda przetwarzania danych, polegająca na usuwaniu nieistotnych cech z zestawu danych.
- d. To proces optymalizacji hiperparametrów modelu w celu uzyskania lepszej dokładności.

Pytanie 4 | Nie udzielono odpowiedzi Punkty maks.: 1,00 Oflaguj pytanie

Czym jest przeuczenie (**overfitting**)?

- a. To zjawisko, w którym model jest zbyt dobrze dopasowany do danych treningowych, co prowadzi do niskiej skuteczności na nowych danych.
- b. To sytuacja, w której model jest niedostosowany do danych, przez co osiąga niską skuteczność zarówno na zbiorze treningowym, jak i testowym.
- c. To metoda przetwarzania danych, polegająca na usuwaniu nieistotnych cech z zestawu danych.
- d. To proces optymalizacji hiperparametrów modelu w celu uzyskania lepszej dokładności.

brak odp A

Pytanie 1

Poprawnie

Punkty: 1,00 z 1,00

Zjawiska nadmiernego dopasowania (ang. **overfitting**) drzewa do danych uczących można unikać poprzez:

Wybierz jedną odpowiedź:

- a. uwzględnianie przykładów o niezdefiniowanych wartościach atrybutów
- b. transformacje drzewa na zbiór reguł
- c. usuwanie niektórych poddrzew i w rezultacie redukcję rozmiarów drzewa ✓
- d. użycie zmodyfikowanej miary informacyjnej (ang. pre-misinformation) zamiast entropii

Poprawna odpowiedź to: usuwanie niektórych poddrzew i w rezultacie redukcję rozmiarów drzewa

8. Różnica pomiędzy training error i generalization error.

BRAK

9. Wyznaczenie optymistycznej i pesymistycznej estymaty przy błędzie uogólnienia.

Wyznacz pesymistyczną estymatę dla zadanych danych. Błąd uogólnienia wynosi $8/30$, parametr Omega jest równy 0.5, a węzeł ma 4 liście.

Odpowiedź: 0.33

**Pytanie 4** | Niepoprawnie Punkty: 0,00 z 1,00 [Oflaguj pytanie](#)

Wyznacz pesymistyczną estymatę dla zadanych danych. Błąd uogólnienia wynosi 0.25, parametr Omega jest równy 0.5, a węzeł ma 10 liści.

Odpowiedź: 0.33



10. Na czym polega cross-validation?

Pytanie 3

Poprawnie

Punkty: 1,00 z 1,00

Na czym polega walidacja krzyżowa (ang. **cross-validation**)?

- a. Na podziale zbioru danych na części, a następnie trenowaniu i testowaniu modelu na tych samych danych.
- b. Na trenowaniu modelu tylko na jednej części zbioru danych, a następnie testowaniu na całym zbiorze danych.
- c. Na wykorzystywaniu wyłącznie jednego zestawu danych treningowych i jednego testowego bez podziału na podzbiory.
- d. Na podziale zbioru danych na kilka podzbiorów, gdzie model jest trenowany na jednych podzbiorach i testowany na innych. ✓

Pytanie 3 | Nie udzielono odpowiedzi Punkty maks.: 1,00 [Oflaguj pytanie](#)

Na czym polega walidacja krzyżowa (ang. **cross-validation**)?

- a. To metoda zwiększania liczby cech w zestawie danych poprzez generowanie nowych kombinacji.
- b. To proces standaryzacji danych przed trenowaniem modelu.
- c. To sposób optymalizacji hiperparametrów modelu za pomocą algorytmów genetycznych.
- d. To technika oceny modelu polegająca na podziale danych na kilka części, gdzie model jest testowany na jednej części danych uczonej na pozostałych.

[Oznacz mój wybór](#)

brak odp GIT D

Pytanie 3

Poprawnie

Punkty: 1,00 z 1,00

Technikę o nazwaną oceną krzyżową (ang. k fold cross validation) do estymacji miar oceny zdolności predykcyjnych klasyfikatora stosuje się, gdy :

Wybierz jedną odpowiedź:

- a. liczba przykładów jest większa niż rozmiar tzw. małej statystycznej próby tj. 33 przykłady
- b. liczba przykładów uczących jest większa niż 100 a mniejsza niż dziesiątki tysięcy ✓
- c. liczba atrybutów jest równa k
- d. liczba przykładów jest rzędu wielu tysięcy przykładów

Pytanie 11

Poprawnie

Punkty: 1,00 z 1,00

Ille razy klasyfikowany jest każdy przykład (jako testowy przez klasyfikator) w procedurze oceny krzyżowej (ang. k fold cross validation)?

Wybierz jedną odpowiedź:

- a. Dokładnie jeden raz. ✓
- b. Przynajmniej raz i nie więcej niż $n - 1$, gdzie n to liczba części na jakie jest podzielony zbiór uczący.
- c. Jest to zależne od liczby części (ang. podziałów), na jakie został losowo podzielony zbiór uczący?
- d. metoda oceny jest techniką wielu losowych podziałów, więc trudno to ustalić
- e. n razy, gdzie n jest zdefiniowane jak w punkcie c.

11. Pre- i post-pruning - na czym polega?

Tzw. „Cost-complexity approach” w indukcji drzew jest wykorzystywane do:

- a. Uczenia się oszczędnego z uwagi na zużycie czasu obliczeń
- b. Wprowadzenie kosztów użycia poszczególnych atrybutów do korekcji miary entropii
- c. Uwzględnienia funkcji kosztu pomyłek w obliczeniach warunków podziału w węzłach
- d. Wykonania redukcji rozmiarów drzewa z sumą ważoną kryteriów minimalizacji błędu klasyfikacji oraz rozmiaru drzewa ✓

1. Jak działa algorytm k-nn?

Pytanie 5

Poprawnie

Punkty: 1,00 z 1,00

Jak działa algorytm k-nn?

- a. Algorytm k-nn klasyfikuje punkty danych za pomocą gradientowego spadku do minimalizacji funkcji kosztu.
- b. Algorytm k-nn szuka najbliższych sąsiadów nowego punktu w zbiorze danych, a następnie przypisuje mu klasę na podstawie większości klas wśród sąsiadów. ✓
- c. Algorytm k-nn tworzy liniowy model, aby przewidywać wartości na podstawie współczynników regresji wyliczonych z danych treningowych.
- d. Algorytm k-nn wykonuje podział danych na grupy na podstawie odległości od środkowych punktów (centroidów), aby zminimalizować różnice w obrębie każdej grupy.

Jak działa algorytm k-nn?

- a. Algorytm k-nn przypisuje nowy punkt do tej klasy, do której należy większość z jego najbliższych sąsiadów.
- b. Algorytm k-nn wykorzystuje funkcje aktywacji do określania wag neuronów w modelu.
- c. Algorytm k-nn oblicza liniową zależność między zmiennymi, aby przewidzieć wartości ciągłe na podstawie punktów treningowych.
- d. Algorytm **k-nn** tworzy centralne grupy punktów danych, a następnie przypisuje nowe punkty do najbliższego centroidu.

brak odp A

2. Na jakie elementy można mieć wpływ w działaniu algorytmu k-nn?

Pytanie 6

Poprawnie

Punkty: 1,00 z 1,00

Na jakie elementy można mieć wpływ w działaniu algorytmu k-nn?

- a. Sposób transformacji danych i liczba warstw neuronowych.
- b. Liczba sąsiadów, metoda obliczania odległości
- c. Liczba klastrów, metoda inicjalizacji centroid
- d. Wielkość zbioru testowego i sposób trenowania modelu.

3. Zalety i wady k-nn

Pytanie 6 | Częściowo poprawnie Punkty: 0,75 z 1,00

Zaznacz wszystkie wady algorytmu k-nn.

- a. Ma problemy z wartościami odstającymi (*ang. outliers*)
- b. Zależny od początkowych wartości
- c. Działa źle dla małych zbiorów danych
- d. Nie radzi sobie z wartościami nominalnymi
- e. Czas klasyfikacji zależny od rozmiaru zbioru danych ✓
- f. Wymaga ręcznego dobrania hiperparametru k ✓
- g. Duże zużycie pamięci ✓

brak odp imo jeszcze dodatkowo D

Pytanie 7

Częściowo poprawne

Punkty: 0,17 z 1,00

Zaznacz wszystkie **wady** algorytmu k-nn.

- a. Duże zużycie pamięci
- b. Nie radzi sobie z wartościami nominalnymi
- c. Ma problemy z wartościami odstającymi (*ang. outliers*) ✗
- d. Działa źle dla małych zbiorów danych
- e. Wymaga ręcznego dobrania hiperparametru k ✓
- f. Czas klasyfikacji zależny od rozmiaru zbioru danych ✓
- g. Zależny od początkowych wartości

brak odp A, B, E, F

Zaznacz wszystkie zalety algorytmu k-nn.

- a. Intuicyjne przypisanie klas ✓
- b. Skaluje się dla dużych zbiorów danych
- c. Bardzo dobrze radzi sobie z wartościami nominalnymi
- d. Odporny na szum w danych ✓
- e. Niewielkie zużycie pamięci
- f. Łatwy w implementacji ✓

4. Radzenie sobie z wartościami nominalnymi

Pytanie 8

Poprawnie

Punkty: 1,00 z 1,00

Jak poradzić sobie z wartościami nominalnymi, kiedy korzystamy z algorytmu k-nn?

- a. Usunąć kolumny z wartościami nominalnymi, aby uprościć model.
- b. Zastąpić wartości nominalne średnią arytmetyczną wszystkich wartości liczbowych w danych.
- c. Zastosować kodowanie wartości nominalnych, takie jak One-Hot Encoding, aby przekształcić je na wartości liczbowe. ✓
- d. Zignorować wartości nominalne, ponieważ k-nn działa tylko na danych liczbowych.

5. Jak wybór różnych wartości k może mieć wpływ na wynik (k=1, 2, ...,n)

Pytanie 9

Poprawnie

Punkty: 1,00 z 1,00

Jak wybór różnych wartości k może mieć wpływ na wynik algorytmu k-nn?

- a. Wartość k nie wpływa na wynik, jeśli używana jest standardowa metryka odległości.
- b. Im większa wartość k, tym większa dokładność modelu, niezależnie od rodzaju danych.
- c. Małe wartości k mogą sprawić, że model będzie bardziej podatny na wartości odstające, podczas gdy większe wartości k mogą prowadzić do bardziej ogólnionych decyzji. ✓
- d. Zwiększenie wartości k sprawia, że algorytm staje się bardziej losowy i mniej dokładny.

Jak wybór różnych wartości k może mieć wpływ na wynik algorytmu k-nn?

- a. Mniejsze wartości k mogą prowadzić do bardziej szczegółowych, ale mniej stabilnych klasyfikacji, natomiast większe wartości k mogą prowadzić do bardziej ogólnych, ale stabilniejszych wyników ✓
- b. Wartość k nie wpływa znacząco na wynik, ponieważ algorytm k-nn zawsze wybiera najbliższego sąsiada
- c. Większe wartości k sprawiają, że algorytm k-nn staje się bardziej czuły na wartości odstające w danych
- d. Im mniejsza wartość k, tym algorytm działa szybciej, ale kosztem utraty dokładności

Pytanie 16

Poprawnie

Punkty: 1,00 z 1,00

Edytowana wersja klasyfikatora k-NN:

Wybierz jedną odpowiedź:

- a. dostosuje wartość k na zbiorze walidującym
- b. polega na wyborze przykładów uczących do tzw. concept description ✓
- c. pozwala na selekcję atrybutów z wykorzystaniem podejścia wrapper
- d. modyfikuje dynamicznie miarę odległości

1. Przekleństwo wymiarowości

Pytanie 10 | Poprawnie Punkty: 1,00 z 1,00 Oflaguj pytanie

Czym jest **kłata wielowymiarowości** (ang. curse of dimensionality)?

- a. To problem wynikający z liniowej zależności między zmiennymi, co powoduje, że model jest mniej dokładny.
- b. To zjawisko, w którym wzrost liczby wymiarów danych powoduje, że dane stają się coraz rzadsze, co utrudnia ich efektywną analizę.
- c. To technika redukcji liczby zmiennych w zestawie danych w celu optymalizacji działania algorytmu.
- d. To sytuacja, w której model traci dokładność z powodu nadmiernego dopasowania do danych treningowych.

Pytanie 11

Niepoprawnie

Punkty: 0,00 z 1,00

Czym jest **kłata wielowymiarowości** (ang. curse of dimensionality)?

- a. To efekt, w którym złożoność modelu wzrasta w miarę zwiększania liczby warstw w sieci neuronowej. ✗
- b. To problem, w którym algorytm k-nn przestaje działać skutecznie, gdy używa się dużych wartości k.
- c. To sytuacja, w której zwiększenie liczby wymiarów danych powoduje, że większość punktów znajduje się daleko od siebie, co utrudnia analizę i klasyfikację.
- d. To proces automatycznego usuwania korelacji między zmiennymi w celu zwiększenia dokładności modelu.

brak odp C

Pojęcie "**przekleństwo wymiarowości**" mówi o:

- a. wzroście złożoności obliczeniowej wraz ze wzrostem liczby atrybutów
- b. żadna z powyższych odpowiedzi nie jest poprawna
- c. konieczności wykładniczego wzrostu liczby przykładów uczących przy rosncej liczbie atrybutów (aby utrzymać jakość klasyfikacji) ✓
- d. spadku jakości klasyfikacji za każdym razem po dodaniu kolejnego atrybutu

3. Czym jest grupowanie?

BRAK

4. Jak działa algorytm?

Pytanie 10

Poprawnie

Punkty: 1,00 z 1,00

Jak działa algorytm k-means?

- a. Algorytm k-means klasyfikuje nowe punkty, przypisując je do klasy większości ich najbliższych sąsiadów.
- b. Algorytm k-means oblicza liniowy model, aby najlepiej dopasować się do danych.
- c. Algorytm k-means dzieli dane na k grup, umieszczając punkty w grupach na podstawie ich odległości od centralnych punktów (centroidów), które są aktualizowane iteracyjnie.

Pytanie 9 | Poprawnie Punkty: 1,00 z 1,00  Oflaguj pytanie

Jak działa algorytm k-means?

- a. Algorytm k-means klasyfikuje dane, bazując na ich najbliższych sąsiadach, wybierając klasę większości sąsiadów.
- b. Algorytm k-means korzysta z gradientu spadku, aby optymalizować funkcję kosztu i dopasować model do danych.
- c. Algorytm k-means przewiduje zależności liniowe między zmiennymi, co pozwala na predykcję wartości liczbowych.
- d. Algorytm k-means przydziela punkty danych do klastrów na podstawie odległości od centroidów, które są optymalizowane iteracyjnie.

Pytanie 9

Poprawnie

Punkty: 1,00 z 1,00

Które ze zdań jest prawdziwe w stosunku do algorytmu k-srednich

- a. Przy wykorzystaniu odległości euklidesowej ma ukierunkowanie do tworzenia kulistych kształtów skupisk ✓
- b. Wykorzystuje rzeczywiste obserwacje (tzw. medoidy) do reprezentacji skupienia
- c. Iteracyjnie próbuje minimalizować miarę zmienności wewnętrz-skupieniowej ✓
- d. Sam algorytm ustala liczbę skupień

5. Jak wybieramy początkowe centroidy?

BRAK

6. Obliczenie jednej iteracji algorytmu

BRAK

7. Jakie są warunki stopu algorytmu?

Podaj trzy warunki stopu algorytmu k-means.

Tekst odpowiedzi Pytanie 11

1. Centroidy prawie się nie zmieniają
 2. Dobrane punkty do klastrów się nie zmieniają
 3. Osiągnięto maksymalną liczbę iteracji
-
2. Warunkiem zatrzymania algorytmu k-srednich jest: Wybierz jedną odpowiedź:
- a) zakończenie dyskretyzacji atrybutów liczbowych
 - b) utworzenie pełnego drzewa skupień
 - c) optymalizacja wartości k - tzn. jest obliczenie
 - d) Osiągnięcie stabilizacji zawartości przykładów wewnętrz skupień

8. Zalety i wady algorytmu

Pytanie 13

Częściowo poprawnie

Punkty: 0,80 z 1,00

Zaznacz wszystkie **zalety algorytmu k-means**.

- a. Niezależny od wartości początkowych
- b. Uogólnia się na klastry o różnych kształtach i rozmiarach, takie jak klastry eliptyczne. ✓
- c. Skaluje się dla dużych zbiorów danych. ✓
- d. Nie wymaga wybierania wartości k manualnie
- e. Relatywnie prosty w implementacji. ✓
- f. Łatwo adaptuje się do nowych danych. ✓
- g. Odporny na problemy związane z wartościami odstającymi (*ang. outliers*)
- h. Gwarantuje zbieżność.

brak odp jeszcze H

Pytanie 14

Nie udzielono odpowiedzi Punkty maks.: 1,00

[Oflaguj pytanie](#)

Zaznacz wszystkie **zalety algorytmu k-means**.

- a. Relatywnie prosty w implementacji.
- b. Gwarantuje zbieżność.
- c. Odporny na problemy związane z wartościami odstającymi (*ang. outliers*)
- d. Nie wymaga wybierania wartości k manualnie
- e. Łatwo adaptuje się do nowych danych.
- f. Skaluje się dla dużych zbiorów danych.
- g. Niezależny od wartości początkowych
- h. Radzi sobie z klastrami o różnych kształtach i rozmiarach.

brak odp A, B, E, F

Pytanie 12 | Poprawnie Punkty: 1,00 z 1,00 [Otaguj pytanie](#)

Zaznacz wszystkie wady algorytmu k-means.

- a. Zależny od wartości początkowych. ✓
- b. Wymaga manualnego wybrania wartości k. ✓
- c. Nie gwarantuje zbliżności.
- d. Ma problemy z radzeniem sobie z wartościami odstającymi. ✓
- e. Ciężki w implementacji.
- f. Nie skaluje się dobrze dla dużych zbiorów danych.

9. Selekcja atrybutów w problemach nienadzorowanych

12. W przypadku uczenia **nienadzorowanego** (np. grupowanie) do selekcji atrybutów NIE można użyć:

- a) usuwania atrybutów skorelowanych
- b) usuwania atrybutów z niską wariancją
- c) tworzenia nowych atrybutów (np. metodą PCA)
- d) usuwania atrybutów z niską wartością information gain**

1. Na czym polega regresja liniowa

Pytanie **14**

Poprawnie

Punkty: 1,00 z 1,00

Na czym polega **regresja liniowa**?

- a. Regresja liniowa to algorytm grupujący dane na podstawie ich odległości od centroidów, które są iteracyjnie optymalizowane.
- b. Regresja liniowa to metoda przewidywania wartości ciągłych na podstawie liniowej zależności między jedną lub kilkoma zmiennymi niezależnymi, a zmienną zależną. ✓
- c. Regresja liniowa to technika klasyfikacyjna, która przypisuje nowe punkty do klasy większości ich najbliższych sąsiadów.
- d. Regresja liniowa to algorytm obliczający odległości między punktami w celu tworzenia klastrów.

Pytanie 13 | Poprawnie Punkty: 1,00 z 1,00 [oflaguj pytanie](#)

Na czym polega regresja liniowa?

- a. Regresja liniowa to algorytm obliczający odległości między punktami w celu tworzenia klastrów.
- b. Regresja liniowa to algorytm grupujący dane na podstawie ich odległości od centroidów, które są iteracyjnie optymalizowane.
- c. Regresja liniowa to technika klasyfikacyjna, która przypisuje nowe punkty do klasy większości ich najbliższych sąsiadów.
- d. Regresja liniowa to metoda przewidywania wartości ciągły na podstawie liniowej zależności między jedną lub kilkoma zmiennymi niezależnymi, a zmienną zależną.

2. Różnica między regresją wielomianową a wieloraką.

BRAK

3. Czym jest obserwacja odstająca

Pytanie 16

Poprawnie

Punkty: 1,00 z 1,00

Czym jest obserwacja odstająca?

- a. To zbiór danych o wysokiej korelacji z innymi zmiennymi niezależnymi.
- b. To wartość średnia, która odzwierciedla centralną tendencję w zbiorze danych.
- c. To punkt danych, który zawsze jest ignorowany w modelowaniu statystycznym.
- d. To nietypowy punkt danych, który wyraźnie odbiega od pozostałych i może wpływać na wyniki analizy statystycznej lub modelowania.

Pytanie 16 | Poprawnie Punkty: 1,00 z 1,00 Oflaguj pytanie

Czym jest obserwacja **odstająca**?

- a. To zbiór danych o wysokiej korelacji z innymi zmiennymi niezależnymi.
- b. To punkt danych, który zawsze jest ignorowany w modelowaniu statystycznym.
- c. To wartość średnia, która odzwierciedla centralną tendencję w zbiorze danych.
- d. To nietypowy punkt danych, który wyraźnie odbiega od pozostałych i może wpływać na wyniki analizy statystycznej lub modelowania.

4. W jaki sposób można uznać obserwację za odstającą.

BRAK

5. Czym jest regresja logistyczna

Pytanie 14 | Poprawnie Punkty: 1,00 z 1,00 Oflaguj pytanie

Czym jest regresja logistyczna?

- a. Regresja **logistyczna** to metoda klasyfikacji danych na podstawie wyznaczenia hiperplaszczyzny oddzielającej klasy.
- b. Regresja logistyczna to metoda tworzenia klastrów danych na podstawie iteracyjnej optymalizacji odległości między punktami.
- c. Regresja logistyczna to technika przewidywania wartości ciągłych za pomocą liniowego dopasowania danych.
- d. Regresja logistyczna polega na wykorzystaniu funkcji logistycznej do modelowania zależności między zmiennymi niezależnymi, a zmienną zależną reprezentującą klasy.

6. Oblicz wartości odds i logit na podstawie danego prawdopodobieństwa.

Pytanie 15 | Zakończone Punkty maks.: 2,00 [Oflaguj pytanie](#)

Rzucamy symetryczną sześcienną kostką. Sukces definiujemy jako wyrzucenie liczby większej niż 4.

Jakie są wartości prawdopodobieństwa sukcesu (**p**), **odds**, i **logit**? Wyraź logit w formie $\ln(a/b)$ gdzie a i b to liczby całkowite.

Tekst odpowiedzi Pytanie 15

$$p = 2/6$$

$$\text{odds} = p/(1 - p) = (2/6)/(1 - 4/6) = 1/2$$

$$\text{logit} = \ln(1/2)$$

