Eksploracja danych
Lista nr 2
1 IV 2025

Zadanie 1 Dyskretyzacja (przedziałowanie) cech ciągłych

a) **Dane:** `iris` (R-pakiet `datasets`).

- Zbiór danych zawiera wyniki pomiarów uzyskanych dla trzech gatunków irysów (tj. *setosa*, *versicolor* i *virginica*) i został udostępniony przez Ronalda Fishera w roku 1936.
- Pomiary dotyczą długości oraz szerokości dwóch różnych części kwiatu – działki kielicha (ang. sepal) oraz płatka (ang. petal).

b) **Wybór cech**
Wykorzystując wybrane narzędzia analizy opisowej, przeanalizuj krótko własności poszczególnych cech i wybierz: jedną cechę o najlepszej i jedną o najgorszej zdolności dyskryminacyjnej (tzn. zdolności do separacji klas/gatunków).

c) **Porównanie nienadzorowanych metod dyskretyzacji**
Dla wybranych w punkcie b) cech zastosuj i porównaj wyniki uzyskane dla różnych metod dyskretyzacji nienadzorowanej (ang. *unsupervised discretization*), przyjmując, że liczba kategorii jest równa liczbie klas (tzn. $K = 3$). W porównaniach uwzględnij algorytmy: *equal width*, *equal frequency*, *k-means clustering* oraz dyskretyzację na bazie przedziałów zadanych przez użytkownika. (*Wskazówka*: można wykorzystać funkcję `discretize()` z R-pakietu *arules*).
Wykorzystując porównanie z rzeczywistymi etykietkami klas (zmienna `Species`), postaraj się rozstrzygnąć, który algorytm dyskretyzacji okazał się najbardziej skuteczny. Czy wyniki otrzymane dla „najlepszej" cechy różnią się istotnie od wyników dla „najgorszej" cechy?

Zadanie 2 Analiza składowych głównych (Principal Component Analysis (PCA))

a) **Dane:** City Quality of Life Dataset (plik `uaScoresDataFrame.csv`, źródło: Kaggle/Teleport.org)

- Zbiór danych zawiera wskaźniki opisujące jakość życia w wybranych miastach, w tym m.in. takie charakterystyki jak: warunki mieszkaniowe, koszty utrzymania, bezpieczeństwo, opieka zdrowotna, edukacja itp.
- Wszystkie cechy ilościowe przyjmują wartości w zakresie 0-10 (większa wartość oznacza lepszy wynik).
- Dodatkowo mamy informację nt. lokalizacji danego miasta (zmienne `UA_Continent` i `UA_Country`).

b) **Przygotowanie danych**
Wczytaj dane do przestrzeni roboczej R'a i zapoznaj się z ich podstawowymi własnościami. Następnie wybierz podzbiór zawierający wyłącznie cechy ilościowe. Porównaj zmienność (wariancję) poszczególnych cech i spróbuj rozstrzygnąć czy konieczne jest zastosowanie standaryzacji (*Wskazówka:* można w tym celu np. wykorzystać wykresy pudełkowe).

c) **Wyznaczenie składowych głównych**
Wyznacz składowe główne i porównaj ich rozrzut wykorzystując wykresy pudełkowe. *Wskazówka:* do wyznaczenia składowych można wykorzystać funkcję `prcomp()` lub `princomp()`. Analizując wektory ładunków (loadings) dla kilku pierwszych składowych (np. PC1, PC2 i PC3), sprawdź, które zmienne mają największy wkład (największą wagę). Spróbuj zinterpretować otrzymane składowe.

d) **Zmienność odpowiadająca poszczególnym składowym**
Zbadaj, jaki procent wyjaśnionej wariancji (zmienności) odpowiada poszczególnym składowym głównym. W szczególności, odpowiedz na pytanie, ile składowych głównych jest potrzebnych do wyjaśnienia: a) 80% lub b) 90% całkowitej zmienności danych.

e) **Wizualizacja danych wielowymiarowych**
Wykorzystaj wyznaczone składowe główne do wizualizacji danych. Można np. wyznaczyć wykresy rozrzutu 2d (lub 3d) dla pierwszych dwóch (lub trzech) składowych głównych. Co na podstawie skonstruowanych wykresów można wywnioskować nt. podobieństwa poszczególnych obiektów (miast)? Czy obiekty układają się w naturalny sposób w grupy? Na wyznaczonych wykresach rozrzutu zidentyfikuj wybrane miasta, (np. te najbardziej różniące się od pozostałych) i krótko je scharakteryzuj.
Uwaga: W interpretacji wyników mogą przydać się dodatkowe informacje nt. lokalizacji poszczególnych miast (w szczególności zmienne `UA_Continent` i `UA_Country`).

f) **Korelacja zmiennych**
Wykorzystując dwuwykres (ang. biplot), zbadaj czy występuje istotna korelacja między poszczególnymi zmiennymi. Wnioski otrzymane na podstawie analizy dwuwykresu porównaj z wnioskami opartymi na macierzy korelacji (funkcja `cor()`).

g) **Końcowe wnioski**
Podsumuj krótko wyniki analiz z poprzednich punktów. Co ciekawego udało się zaobserwować? Ile składowych potrzebujemy aby otrzymać zadowalającą reprezentację danych? Czy (nie)zastosowanie standaryzacji miało istotny wpływ na otrzymane wyniki i wnioski?

ZADANIE 3 **Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))**

a) **Dane:** `titanic_train` (R-pakiet `titanic`)

- Zbiór danych zawiera wybrane charakterystyki opisujące pasażerów Titanica (w tym m.in. takie zmienne jak: wiek, płeć, miejsce rozpoczęcia podróży czy klasa pasażerska) wraz z informacją czy dana osoba przeżyła katastrofę (zmienna `Survived`).
- Dokładniejszy opis danych: https://www.kaggle.com/c/titanic/data

b) **Przygotowanie danych**
Zapoznaj się z opisem danych i wczytaj je do przestrzeni roboczej R. Sprawdź czy wszystkie typy zmiennych zostały poprawnie przypisane i w razie potrzeby wykonaj odpowiednią konwersję (*Wskazówka:* w tym celu można wykorzystać funkcje `as.factor()`, `as.ordered()` itp.). Następnie usuń z danych zmienne pełniące rolę identyfikatorów pasażerów (takie jak: PassengerId, Name, Ticket i Cabin).
Uwaga: kolumnę `Survived` traktujemy jako zmienną grupującą i nie będziemy jej wykorzystywali do przeprowadzenia redukcji wymiaru, a jedynie na etapie interpretacji wyników.

c) **Redukcja wymiaru na bazie MDS**
Dla przygotowanych w punkcie b) danych wyznacz (z pominięciem zmiennej `Survived`) macierz odmienności (ang. dissimilarity matrix) i zastosuj wybrany wariant skalowania wielowymiarowego (tj. skalowanie metryczne lub niemetryczne). Jako wymiar docelowej przestrzeni przyjmij $d = 2$ lub 3. Zbadaj jakość otrzymanego odwzorowania, wykorzystując w tym celu diagram Sheparda.

d) **Wizualizacja danych**
Konstruując odpowiednie wykresy rozrzutu (2D lub 3D), przedstaw graficznie wyniki redukcji wymiaru. Wykorzystując różne kolory, zaznacz przynależność poszczególnych jednostek do grup odpowiadających wartościom zmiennej `Survived` (tj. `Survived==0` i `Survived==1`). Zinterpretuj otrzymane wyniki. Czy widoczny jest podział obiektów na grupy (skupiska)? Czy i w jakim stopniu jest on zgodny z informacją dot. przeżycia katastrofy? Czy na wykresie widoczne są obserwacje nietypowe (odstające)?
Skonstruuj również analogiczne wykresy rozrzutu (2D lub 3D), zaznaczając tym razem różnymi kolorami (lub symbolami) wartości (poziomy) zmiennej `Sex` oraz `Pclass`. Czy w tym przypadku widoczne na wykresie grupy (skupiska) są powiązane z wartościami tych zmiennych?