Raport nr 2

Emilia Kowal [249716], Jakub Dworzański [249703]

26 kwietnia 2020

Spis treści

1 Krótki opis zagadnienia

2 Opis eksperymentów/analiz

3 Wyniki

4 Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))

4.1 Krótki opis zagadnienia

W tej części będziemy się zajmowali MDS, czyli skalowaniem wielowymiarowym. Zdecydowaliśmy się na zbadanie skalowania niemetrycznego, które jest wariantem MDS.

4.2 Opis eksperymentów/analiz

Będziemy sprawdzali jakość odwzorowania MDS. W tym celu zbadamy, jak zmieniają się wartości funkcji `STRESS` oraz diagramy Shepparda dla różnych wymiarów docelowej przestrzeni. Do badań wykorzystamy zbiór danych, dotyczący pasażerów Titanica.

Tabela 1: Przykładowe dane

Pclass	Sex	Age	sibsp	Parch	Fare	Embarked	Survived
3	M	22	1	0	7.2500	2	N
1	F	38	1	0	71.2833	0	Y
3	F	26	0	0	7.9250	2	Y
1	F	35	1	0	53.1000	2	Y
3	M	35	0	0	8.0500	2	N
3	M	28	0	0	8.4583	1	N

4.3 Wyniki

```
head(titanic.dane)
```

W tabeli 1, przedstawiającej przykładowe dane, możemy zobaczyć, że wśród zmiennych występują nie tylko cechy numeryczne, takie jak wiek(`Age`), ale również cechy kategoryczne, w tym:

- płeć(`Sex`), jako zmienna binarna,
- port(`Fare`), w którym pasażer wszedł na pokład statku, jako zmienna nominalna,
- klasę(`Pclass`), którą podróżował pasażer, jako zmienna uporządkowana.

Ponadto w danych istnieje zmienna grupująca `Survived`, która informuje o tym, czy pasażer przeżył katastrofę.

Do dalszej analizy będziemy wykorzystywać dane bez zmiennej grupującej, aby później na jej podstawie ocenić jakość odwzorowania.

```
titanic.dane.mds <- subset(titanic.dane, select=-c(Survived))
```

W kolejnych krokach, aby dokonać skalowania, będziemy potrzebować odmienności między wektorami cech.

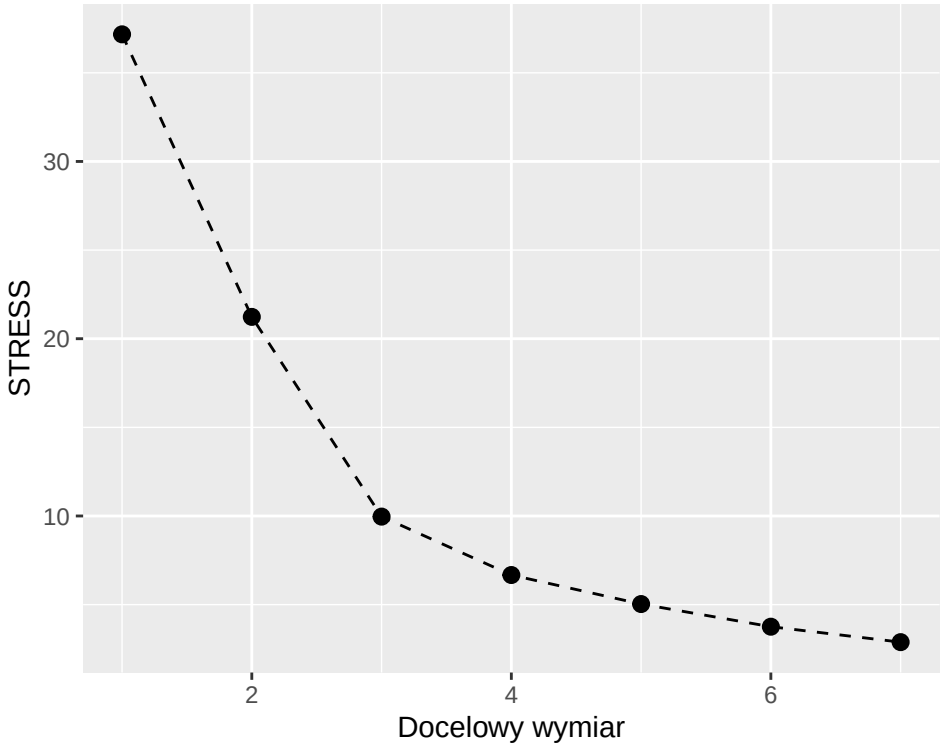
```
macierz.odmiennosci <- as.matrix(daisy(
  titanic.dane.mds,
  stand=TRUE
))
```

Dysponując macierzą odmienności, możemy zbadać jak zmieniają się diagramy Sheppard-da oraz wartość funkcji kryterialnej `STRESS`, wraz ze zmianą wymiaru docelowej przestrzeni. Dlatego teraz wyznaczamy `STRESS` oraz diagram Shepparda dla k=1,2,...,7.

```
badanie.MDS <- function (k.max, macierz.odmiennosci) {
  STRESS <- numeric(k.max)
  wykresy <- list()
  for (k in 1:k.max) {
    mds <- isoMDS(macierz.odmiennosci, k=k)
```

```
    STRESS[k] <- mds$stress
    odleglosci.k <- as.matrix(dist(mds$points, method="euclidean"))
    wykresy[[k]] <- ggplot() + geom_point(aes(x=c(macierz.odmiennosci), y=c(odleglosci.k
  }
  return(list(wykresy=wykresy, STRESS=STRESS))
}
k.max <- length(titanic.dane.mds)
wyniki <- badanie.MDS(k.max, macierz.odmiennosci)
```

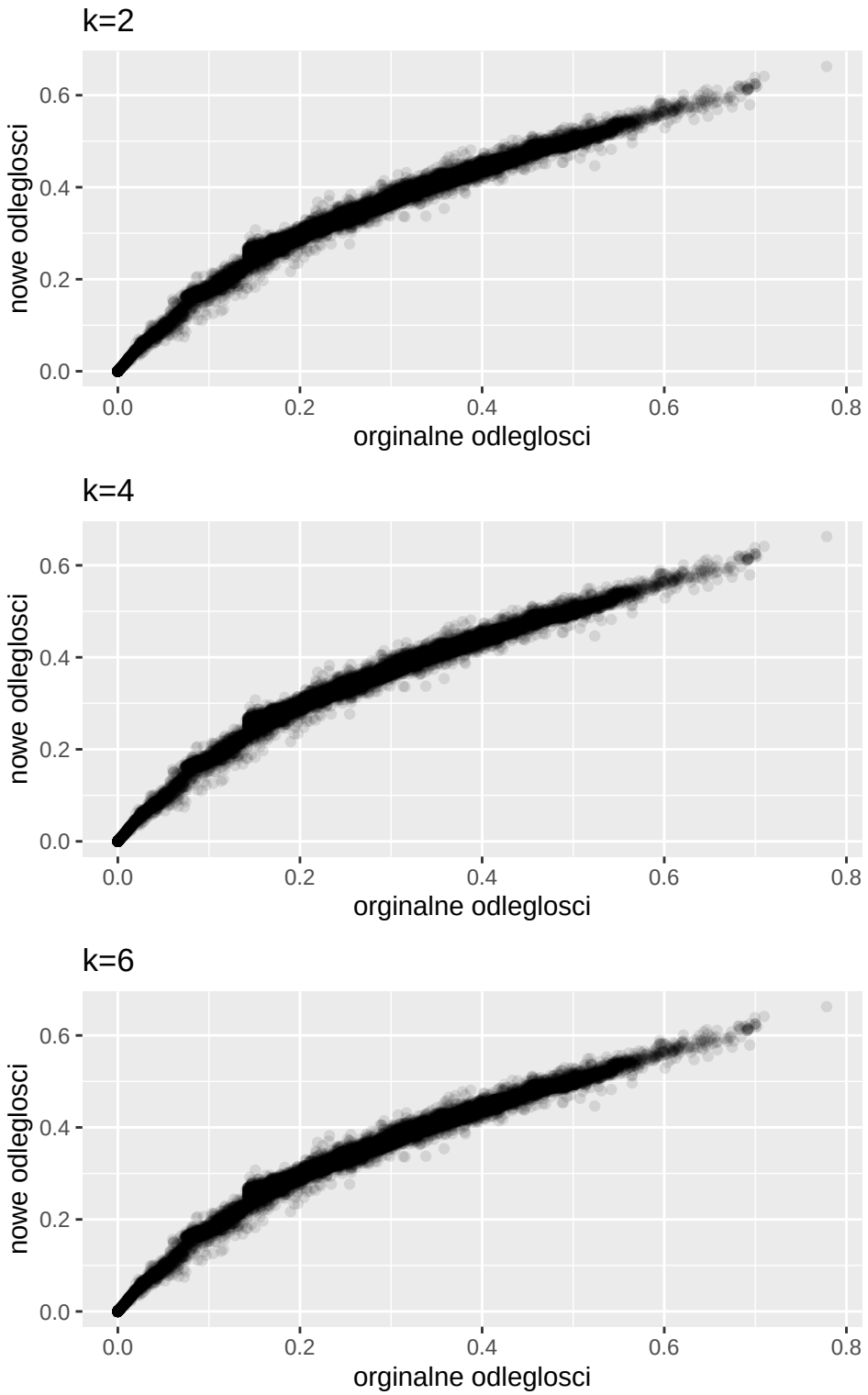
```
ggplot() + geom_line(aes(x=1:k.max, y=wyniki$STRESS))
```



Rysunek 1: Zależność funkcji kryterialnej STRESS od docelowego wymiaru

Na wykresie 1 widzimy, że funkcja `STRESS` maleje z każdym zwiększeniem wymiaru docelowego. Na jedo podstawie, możemy również stwierdzić, że dla w przestrzeni dwuwymiarowej przyjmuje ona stosunkowo duże wartości, co wskazuje na potencjalnie dużą utratę informacji. Widzimy również, że także w przestrzeni trójwymiarowej występuje utrata informacji, jednak jest ona znacznie mniejsza, niż dla $k = 2$. Stąd, możemy przypuszczać, że dzięki wykorzystaniu metod wizualizacji 3D, będziemy mogli poznać strukturę zbioru danych znacznie lepiej, niż na standardowym, dwuwymiarowym wykresie.

```
wykresy <- wyniki$wykresy
ggarrange(plotlist=wykresy[c(2, 4, 6)], nrow=3)
```



Rysunek 2: Diagramy Shepparda

Diagramy Shepparda, które przedstawiają porównanie odległości między orginalną, a docelową przestrzeną, nie przedstawiają istotnych zmian między docelowymi wymiarami. Dlatego też, korzystając przede wszystkim z wykresu funkcji `STRESS`, jako docelowy wymiar przestrzeni, wybrałbym 4, ponieważ wykres zaczyna się w tym miejscu wypłaszczać, a jednocześnie pozwala on na wizualne badanie elementów zbioru poprzez wprowadzenie koloru lub kształtu, reprezentującego czwartą współrzędną w tej przestrzeni.

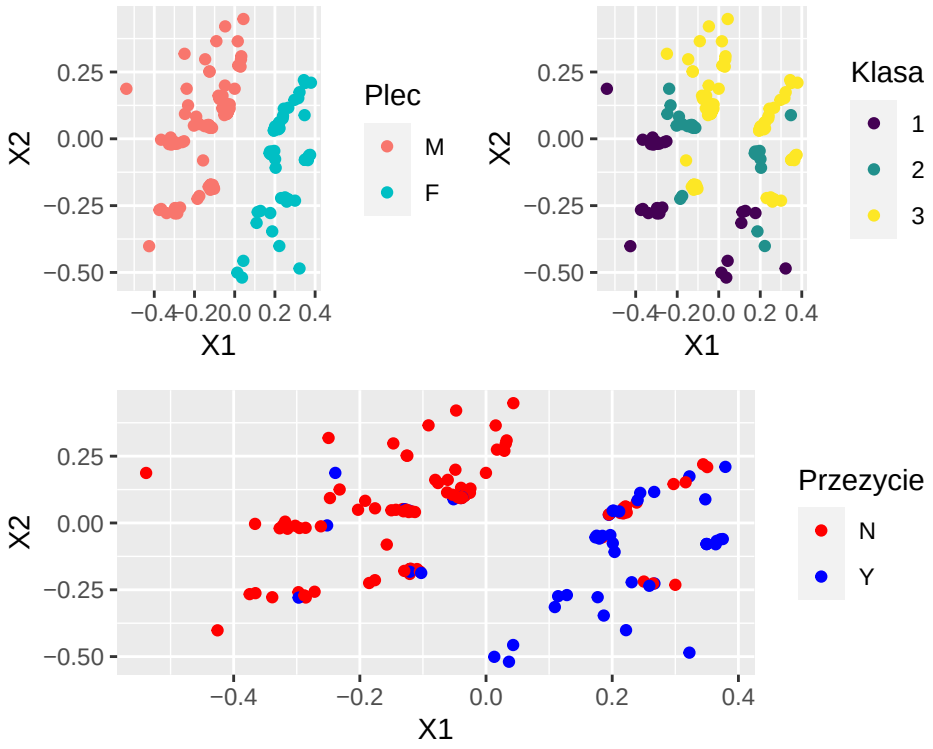
Tabela 2: Tabela kontygencji dla podziału płeć-przeżycie

	N	Y
M	83	13
F	16	38

Teraz spróbujemy ocenić jakość odwzorowań, korzystając z dodatkowej informacji, jaką jest zmienna grupująca `Survived`.

```
mds.k2 <- isoMDS(macierz.odmiennosci, k=2)
reprezentacja.mds.k2 <- data.frame(mds.k2$points)

ggplot(reprezentacja.mds.k2, aes(x=X1, y=X2)) +
  geom_point(aes(col=...))
```



Rysunek 3: Wykres po odwzorowaniu MDS dla $k = 2$ z podzialem na grupy

Widzimy, że po odwzorowaniu można wyróżnić dwie grupy, które rodzielają dane ze względu na płeć pasażera. Grupy te nie rozdzielają jednak danych ze względu na informację dotycząca tego, czy pasażer przeżył. Możemy jednak zauważyć, że istotnie wśród kobiet odsetek przeżycia jest dużo wyższy. Stąd, po sprawdzeniu, jak prezentuje się ich rozkład:

```
table(titanic.dane$Sex, titanic.dane$Survived)
```

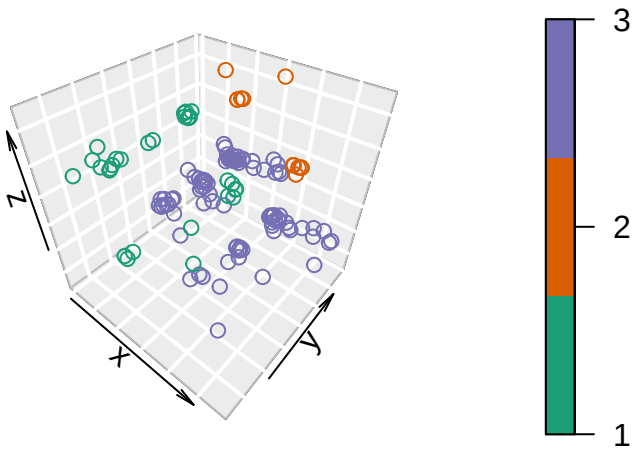
W tabeli 2 widzimy, że faktycznie, przypisując tym dwóm grupom odpowiednią klasę grupującą, otrzymalibyśmy skuteczność na poziomie 71.59%. Stąd możemy powiedzieć, że dane po MDS wciąż zachowują dość dużo informacji.

Ponadto, na wykresie 2 widzimy, że w danych pozostał dość dobry podział ze względu na klasę, którą podróżował pasażer.

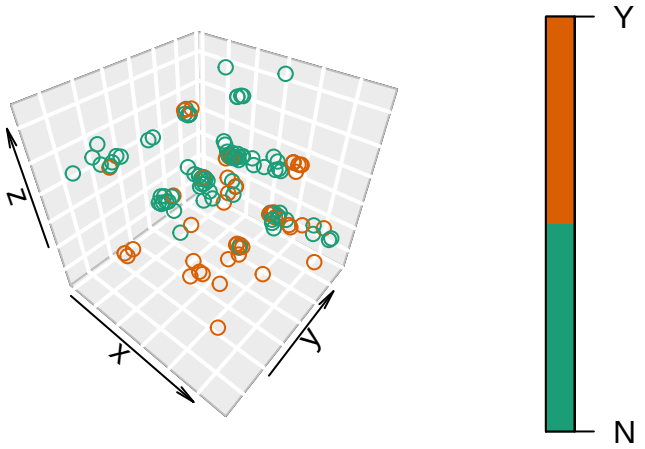
Teraz, możemy sprawdzić, czy przekształcenie MDS do przestrzeni trojwymiarowej będzie zawierało więcej informacji.

```
mds.k3 <- isoMDS(macierz.odmiennosci, k=3)
attach(data.frame(mds.k3$points))
```

```
points3D(X1,X2,X3,...)
```



Rysunek 4: Wykres po odwzorowaniu MDS dla $k = 3$ z podzialem na początek podróży



Rysunek 5: Wykres po odwzorowaniu MDS dla $k = 3$ z podziałem na przeżycie

Na wykresie 4 widzimy, że dodatkowy wymiar dostarcza nam informacji między innymi na temat portu, w którym pasażer wsiadł na statek.

Nie jest to jednak wystarczające do separacji pasażerów, którzy przeżyli, od tych, którzy nie przeżyli (Rys. 5). Mimo tego, redukcja wymiaru zdecydowanie pomogła w wizualizacji i w lepszym zrozumieniu danych.

4.4 Podsumowanie

Na podstawie przeprowadzonego eksperymentu, możemy stwierdzić, że skalowanie wielowymiarowe może być niezwykle pomocne podczas analizy danych. Mimo tego, że nie rozwiązało ono bezpośrednio głównego problemu, który wiąże się z tym zbiorem danych, czyli predykcji przeżycia pasażerów, to pozwoliło na przeanalizowanie danych w 2- i 3- wymiarowych przestrzeniach, co pozwoliło na sformułowanie początkowych hipotez, które mogłyby posłużyć do dalszej analizy.

5 Podsumowanie

Literatura