Wykład 16

Przykład

W trakcie Igrzysk Olimpijskich w Sztokholmie (1912 rok) rozegrano po raz pierwszy w ramach zawodów lekkoatletycznych 10-bój. Konkurencja odbyła się od 13. lipca do 15. lipca. Udział wzięło 29 zawodników, ukończyło konkurencję 12. Na dzisiaj za zwycięzców[1] uznaje się Amerykanina Jima Thorpe i Szweda Hugo Wieslandera.

W skład 10-boju wchodzą następujące konkurencje: bieg na 100 metrów, skok w dal, pchnięcie kulą, skok wzwyż, bieg na 400 metrów, rzut dyskiem, bieg na 110 metrów przez płotki, skok o tyczce, rzut oszczepem i bieg na 1500 metrów. Wyniki przeliczane są (według tabel/wzorów) na punkty, punkty są sumowane i tak powstaje końcowa klasyfikacja.

Wiersze tabeli wyników nazywać będziemy osobnikami, kolumny zaś – zmiennymi. Plik DECA1912.CSV wyniki zawiera zawodów. Zapisano go z użyciem kodowania UTF-8, znakiem rozdzielającym i znakiem dziesiętnym jest przecinek. Zawartość pola jest zawarta w cudzysłowiach, dlatego nie powoduje to konfliktu.

```
> data <- read.csv("deca1912.csv", dec=",", encoding="UTF-8")
> head(data, n=3)
     Imie  Nazwisko kraj punkty m100 wdal  kula wzwyz m400  dysk m110 tyczka
1 Alfreds Alslebens  RUS   5295 12.2 6.27  8.48  1.70 59.0 29.21 19.5    0.0
2   James   Donahue  USA   6784 11.8 6.48  9.67  1.65 51.6 29.95 16.2    3.4
3    Karl      Halt  GER   6683 12.1 6.08 11.12  1.70 54.2 35.46 17.7    2.7
  oszcz m1500 final
1 37.34 308.6    12
2 37.09 284.0     5
3 39.82 302.8     9
> names(data)
 [1] "Imie"     "Nazwisko" "kraj"     "punkty"   "m100"     "wdal"
 [7] "kula"     "wzwyz"    "m400"     "dysk"     "m110"     "tyczka"
[13] "oszcz"    "m1500"    "final"
> nrow(data)
[1] 12
> class(data)
[1] "data.frame"
```

Poniżej kilka podstawowych poleceń służących do manipulacji danymi:

```
> A <- as.matrix(data[,5:14])
> head(A, n=3)
     m100 wdal  kula wzwyz m400  dysk m110 tyczka oszcz m1500
[1,] 12.2 6.27  8.48  1.70 59.0 29.21 19.5    0.0 37.34 308.6
[2,] 11.8 6.48  9.67  1.65 51.6 29.95 16.2    3.4 37.09 284.0
[3,] 12.1 6.08 11.12  1.70 54.2 35.46 17.7    2.7 39.82 302.8
> A[3,"m100"]
m100      12.1
> A[3,]
```

[1] To historia na długą opowieść – ale z innego przedmiotu.

```
  m100   wdal   kula  wzwyz   m400   dysk   m110 tyczka  oszcz  m1500
 12.10   6.08  11.12   1.70  54.20  35.46  17.70   2.70  39.82 302.80
> data$m100
 [1] 12.2 11.8 12.1 11.4 12.3 11.8 11.0 12.3 12.3 11.2 11.5 11.8
> data[3,]$m100   [1] 12.1
> data[3,"m100"]   [1] 12.1
```

Dla macierzy A wyznaczmy macierze AA^T oraz A^TA. Pierwsza z nich ma rozmiar 12×12, druga – 10×10. Wyznaczmy też wartości własne tych macierzy:

```
> AAT <- A %*% t(A)
> dim(AAT)
[1] 12 12
> evAAT <- eigen(AAT)
> names(evAAT)
[1] "values"  "vectors"
> evAAT$values
 [1]  1.100039e+06  4.805804e+02  7.798585e+01  4.420662e+01  6.110277e+00
 [6]  2.284394e+00  2.024427e+00  5.507802e-01  1.770528e-01  6.136206e-03
[11] -1.672044e-13 -1.245011e-11
> ATA <- t(A) %*% A
> dim(ATA)
[1] 10 10
> evATA <- eigen(ATA)
> evATA$values
 [1] 1.100039e+06 4.805804e+02 7.798585e+01 4.420662e+01 6.110277e+00
 [6] 2.284394e+00 2.024427e+00 5.507802e-01 1.770528e-01 6.136206e-03
```

Jak można zauważyć wartości własne (od λ_1 do λ_{10}) są takie same. Dodatkowo: w rozkładzie $A = U\Sigma V^T$ macierze U, V to ortonormalne wektory własne AA^T, A^TA. Niech $\sigma_k = \sqrt{\lambda_k}$. Wiadomo[2], że $Av_k = \sigma_k u_k$.

```
> U <- evAAT$vectors
> V <- evATA$vectors
> dim(U)     [1] 12 12
> dim(V)     [1] 10 10
> max(U %*% t(U) - diag(12))
[1] 3.85976e-16
> max(V %*% t(V) - diag(10))
[1] 6.661338e-16
```

Macierze U, V są zatem ortogonalne. Jeżeli chodzi o związek $Av_k = \sigma_k u_k$:

```
> t(A %*% V[,1])
          [,1]      [,2]      [,3]      [,4]      [,5]     [,6]      [,7]
[1,] -318.6196 -293.4756 -313.1786 -293.2555 -295.1444 -322.681 -293.8773
          [,8]      [,9]     [,10]     [,11]     [,12]
```

[2] Np. z poprzedniego wykładu.

```
[1,] -319.9338 -296.8403 -291.8123 -294.3111 -297.4409
> sqrt(evATA$values[1]) * U[,1]
 [1] -318.6196 -293.4756 -313.1786 -293.2555 -295.1444 -322.6810 -293.8773
 [8] -319.9338 -296.8403 -291.8123 -294.3111 -297.4409
> max(abs(A%*%V) - abs(U[,1:10]%*%Sigma))
[1] 1.659814e-10
```

Zmieńmy dane dotyczące skoku wzwyż, tzn. zapiszmy wynik w centymetrach.

```
> evATA$values
 [1] 1.100039e+06 4.805804e+02 7.798585e+01 4.420662e+01 6.110277e+00
 [6] 2.284394e+00 2.024427e+00 5.507802e-01 1.770528e-01 6.136206e-03
> evBTB$values
 [1] 1.438905e+06 1.296138e+03 3.250227e+02 7.521227e+01 2.868788e+01
 [6] 5.528987e+00 2.049161e+00 7.072035e-01 2.137084e-01 1.376007e-01
> sum(evATA$values)   [1] 1100653
> sum(evBTB$values)   [1] 1440638
```

Zmieniają się wartości własne (chociaż na pierwszy rzut oka nie tak bardzo), zmienia się też suma wartości własnych. Wprawdzie nie jest to kluczowym niedostatkiem – wspomnijmy chociażby fakt iż suma wartości własnych jest równa sumie elementów przekątniowych – pozostaje jednak faktem iż skalowanie (zmiana jednostki pomiarowej) zmienia wartości własne.

Powróćmy teraz do interpretacji macierzy A^TA oraz AA^T. Przypominając interpretację wierszy i kolumn macierzy A jako osobniki i zmienne stwierdzamy, że elementami pierwszej macierzy są iloczyny skalarne zmiennych a drugiej – iloczyny skalarne osobników. Jeżeli wstępnie odejmiemy od kolumn macierzy A ich średnie, to otrzymamy (z dokładnością do liczby osobników m) macierz wariancji-kowariancji zmiennych, ponieważ

$$\left(A^TA\right)_{ij} = (c_i - \bar{c_i})^T (c_j - \bar{c_j}), \text{ gdzie A} = \left[\ c_1 : \ldots : c_n\ \right].$$

Kolejnym krokiem jest podzielenie każdej ze zmiennych przez odchylenie standardowe. Otrzymamy w efekcie macierz korelacji zmiennych, na przekątnej będziemy mieli wartości 1. Dodatkowo jest też

$$\sum_{k=1}^{n} \sigma_k^2 = n.$$

Ostatnia równość umożliwia prostą ocenę skuteczności (efektywności) kolejnych wektorów i wartości własnych. Wielkość σ_k^2/n można uważać za udział k-tego kierunku w łącznym zróżnicowaniu zmiennych.

Zadania

1. W pliku DECA1920.CSV znajdują się wyniki 10-boju z IO 1920 roku.
   (a) Wyznaczyć trzy największe wartości własne $\sigma_1^2, \sigma_2^2, \sigma_3^2$, macierzy $A^T A$.
   (b) Jaki procent sumy wartości własnych stanowią te trzy wartości własne?
   (c) Podać wektory własne u_1, v_1 odpowiadające wartości własnej σ_1^2.
2. Zamienić wyniki skoku wzwyż, skoku w dal i skoku o tyczce na centymetry. Wyznaczyć trzy największe wartości własne $\sigma_1^2, \sigma_2^2, \sigma_3^2$, macierzy $A^T A$. Jaki procent sumy wartości własnych stanowią te trzy wartości własne?
3. Od każdego wyniku odjąć średni wynik danej konkurencji, różnicę podzielić przez odchylenie standardowe tejże konkurencji. Dla otrzymanej macierzy powtórzyć trzy podpunkty z zadania 1.