ALGORYTMICZNA I STATYSTYCZNA ANALIZA DANYCH

18/12/2014

WFAiS UJ, Informatyka Stosowana
II stopień studiów

Eksploracja danych

Analiza danych Samsung

Analiza danych pm25

Eksploracja danych

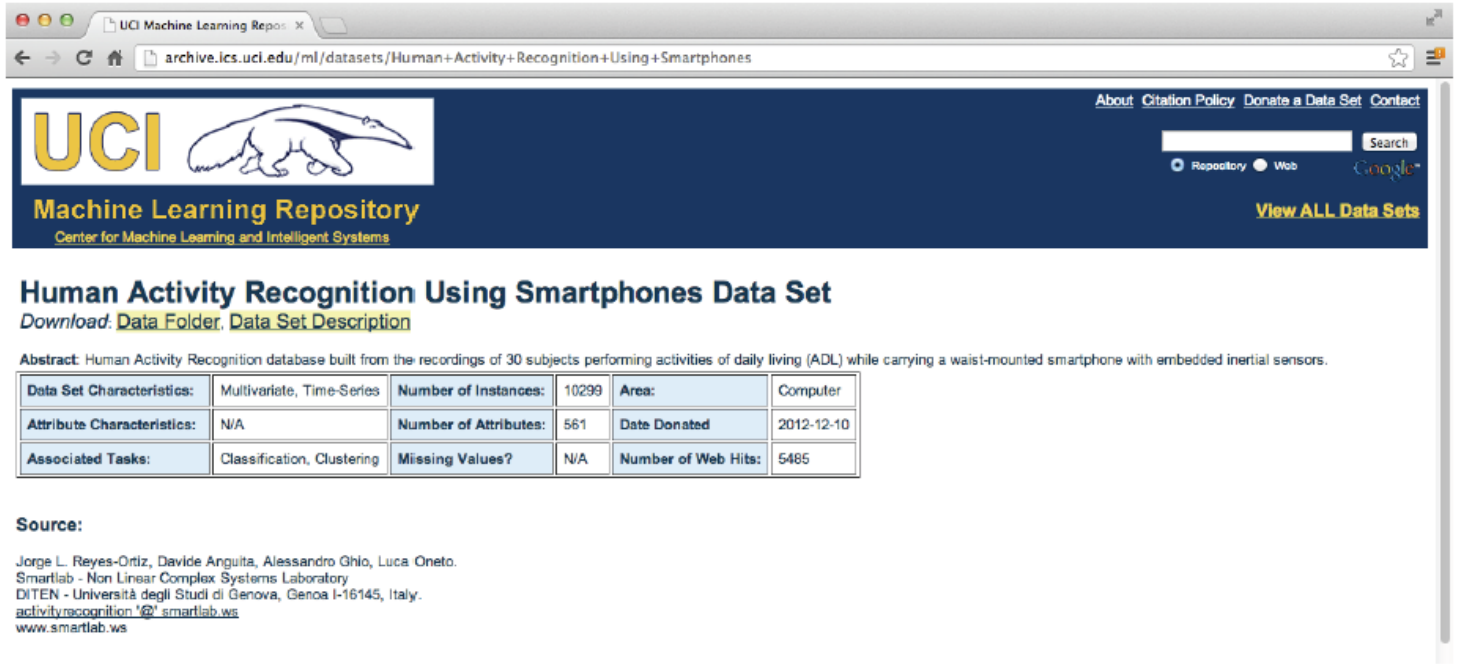
- Taje nam bardzo „zgrubny” obrazek jaka informacje możemy z danych wyciągnąć i na jakie pytania możemy odpowiedzieć.
- Przewiduje poruszanie się ludzi na podstawie danych zarejestrowanych przez ich smartfony.

http://www.samsung.com/global/galaxys3/

W każdym smartfonie jest akcelerometer i żyroskop, tak ze możemy rejestrować Pozycję i prędkość poruszania się

Exploracja danych

Samsung Data



http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

Dane:


```
load("data/samsungData.rda")
names(samsungData)[1:12]
```

Przyspieszenie ciała w kierunku x,y,z.

```
##  [1] "tBodyAcc-mean()-X" "tBodyAcc-mean()-Y" "tBodyAcc-mean()-Z"
##  [4] "tBodyAcc-std()-X"  "tBodyAcc-std()-Y"  "tBodyAcc-std()-Z"
##  [7] "tBodyAcc-mad()-X"  "tBodyAcc-mad()-Y"  "tBodyAcc-mad()-Z"
## [10] "tBodyAcc-max()-X"  "tBodyAcc-max()-Y"  "tBodyAcc-max()-Z"
```

- Czy możemy przewidzieć aktywność w oparciu o te dane?

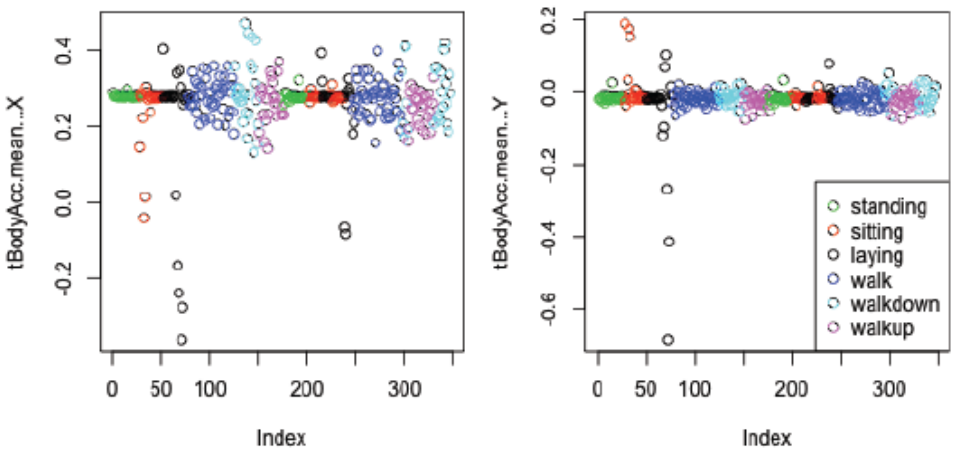
```
table(samsungData$activity)
```

Aktywność: leży, siedzi, stoi, idzie, idzie pod górę, idzie na dół

```
##
##   laying  sitting standing     walk walkdown   walkup
##     1407     1286     1374     1226      986     1073
```

Średnie przyspieszenie


```
par(mfrow = c(1, 2), mar = c(5, 4, 1, 1))
samsungData <- transform(samsungData, activity = factor(activity))
sub1 <- subset(samsungData, subject == 1)
plot(sub1[, 1], col = sub1$activity, ylab = names(sub1)[1])
plot(sub1[, 2], col = sub1$activity, ylab = names(sub1)[2])
legend("bottomright", legend = unique(sub1$activity), col = unique(sub1$activity),
    pch = 1)
```

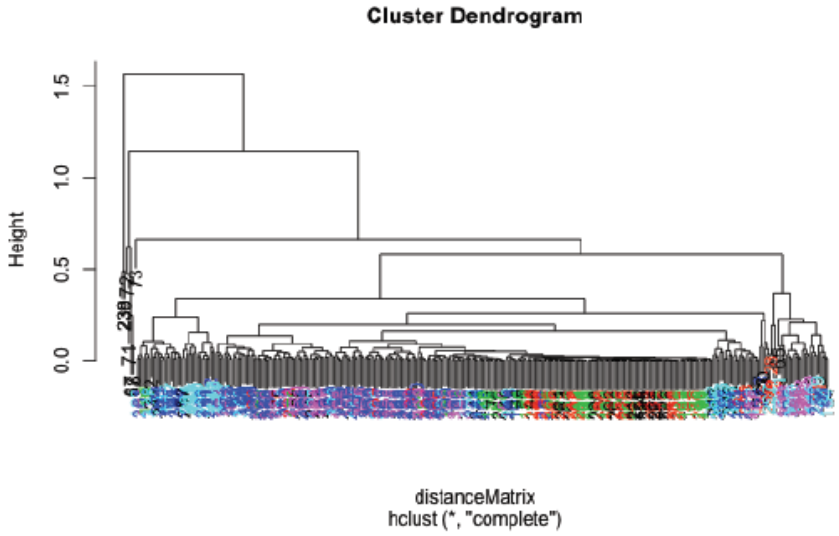


Istotna zmienność dla

- Walk
- Walk down
- Walk up

Klastering na podstawie średniego przyspieszenia

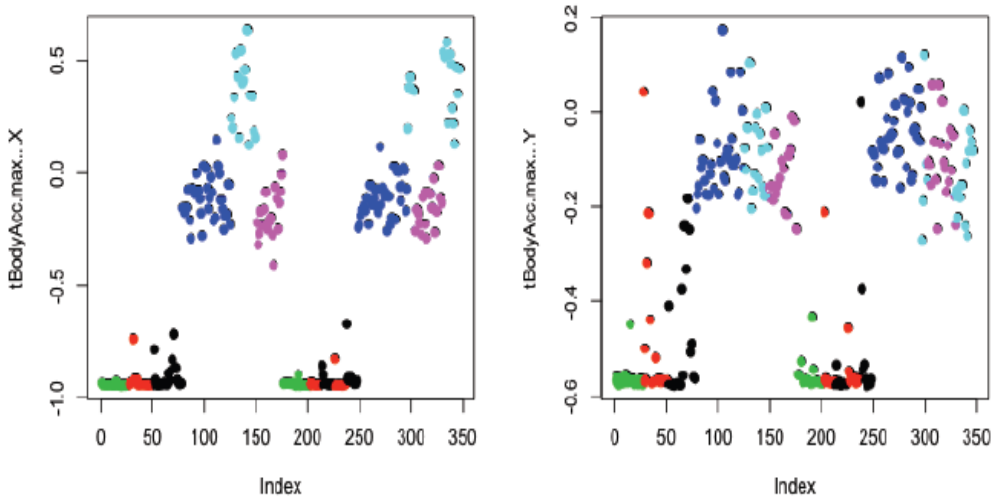

```
source("myplclust.R")
distanceMatrix <- dist(sub1[, 1:3])
hclustering <- hclust(distanceMatrix)
myplclust(hclustering, lab.col = unclass(sub1$activity))
```



Nie widać dobrej separacji klastrów ze względu na Średnią aktywności.
Ta zmienna nie pozwala na dobrą separację.

Klastering na podstawie maksymalnego przyspieszenia

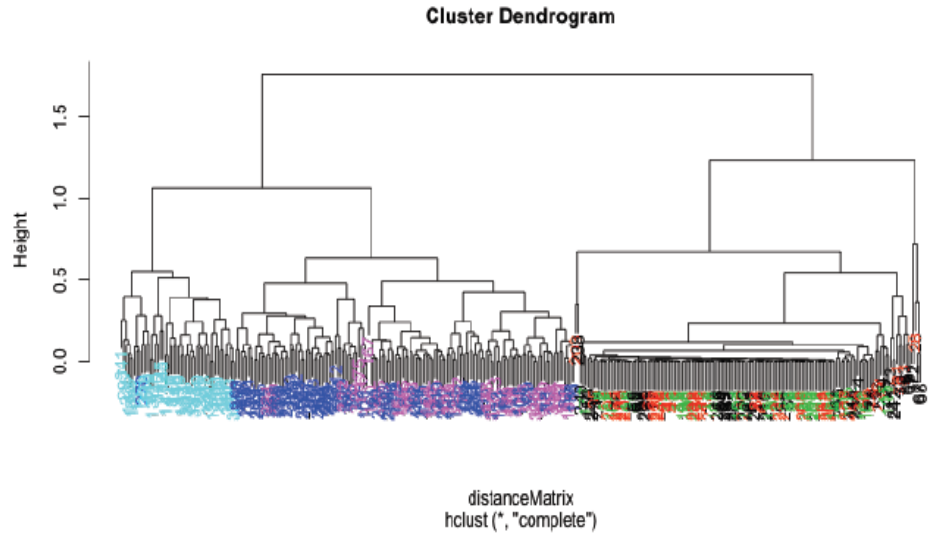

```
par(mfrow = c(1, 2))
plot(sub1[, 10], pch = 19, col = sub1$activity, ylab = names(sub1)[10])
plot(sub1[, 11], pch = 19, col = sub1$activity, ylab = names(sub1)[11])
```



Widać separacje ze względu na maksymalną aktywność.

Klastering na podstawie maksymalnego przyspieszenia

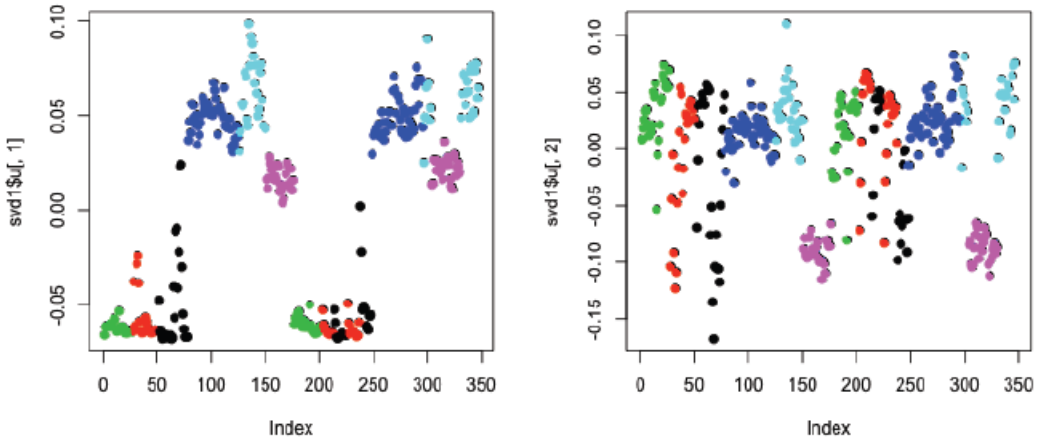

```
source("myplclust.R")
distanceMatrix <- dist(sub1[, 10:12])
hclustering <- hclust(distanceMatrix)
myplclust(hclustering, lab.col = unclass(sub1$activity))
```



Widać separacje ze względu na maksymalną aktywność.
Dwa klastry:
- walk, walkup, walkdown
- laying, standing, sitting

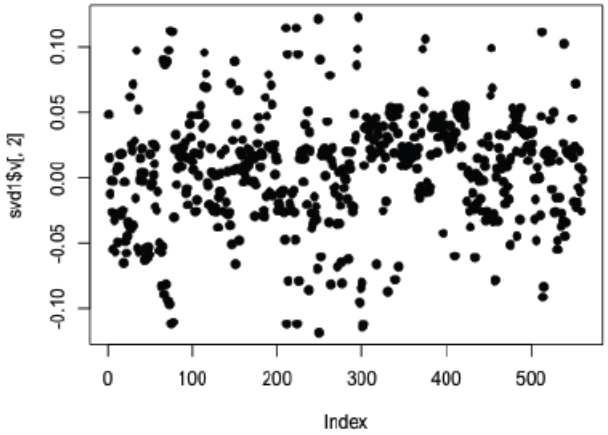
SVD analysis


```
svd1 = svd(scale(sub1[, -c(562, 563)]))
par(mfrow = c(1, 2))
plot(svd1$u[, 1], col = sub1$activity, pch = 19)
plot(svd1$u[, 2], col = sub1$activity, pch = 19)
```



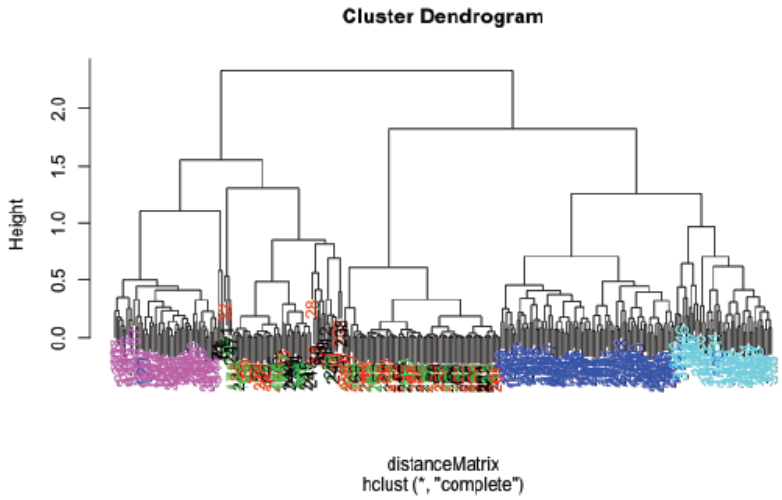
Znajdź „maximum contributor”


```
plot(svd1$v[, 2], pch = 19)
```



Przeprowadź klastering używając tej zmiennej


```
maxContrib <- which.max(svd1$v[, 2])
distanceMatrix <- dist(sub1[, c(10:12, maxContrib)])
hclustering <- hclust(distanceMatrix)
myplclust(hclustering, lab.col = unclass(sub1$activity))
```




```
names(samsungData)[maxContrib]
```

```
## [1] "fBodyAcc.meanFreq...Z"
```

K-means klastering


```
kClust <- kmeans(sub1[, -c(562, 563)], centers = 6)
table(kClust$cluster, sub1$activity)
```

```
##
##      laying sitting standing walk walkdown walkup
##   1       0       0        0   50        1      0
##   2       0       0        0    0       48      0
##   3      27      37       51    0        0      0
##   4       3       0        0    0        0     53
##   5       0       0        0   45        0      0
##   6      20      10        2    0        0      0
```

Też ma problem
Aby rozróżniać
laying
Sitting
standing

K-means klastering (następna próba)


```
kClust <- kmeans(sub1[, -c(562, 563)], centers = 6, nstart = 1)
table(kClust$cluster, sub1$activity)
```

```
##
##      laying sitting standing walk walkdown walkup
##   1       0       0        0    0       49      0
##   2      18      10        2    0        0      0
##   3       0       0        0   95        0      0
##   4      29       0        0    0        0      0
##   5       0      37       51    0        0      0
##   6       3       0        0    0        0     53
```

Też ma problem
Aby rozróżniać
laying
Sitting
standing

K-means klastering (następna próba)


```
kClust <- kmeans(sub1[, -c(562, 563)], centers = 6, nstart = 100)
table(kClust$cluster, sub1$activity)
```

```
##
##      laying sitting standing walk walkdown walkup
##   1      18      10        2    0        0      0
##   2      29       0        0    0        0      0
##   3       0       0        0   95        0      0
##   4       0       0        0    0       49      0
##   5       3       0        0    0        0     53
##   6       0      37       51    0        0      0
```

K-means klastering (następna próba)


```
kClust <- kmeans(sub1[, -c(562, 563)], centers = 6, nstart = 100)
table(kClust$cluster, sub1$activity)
```

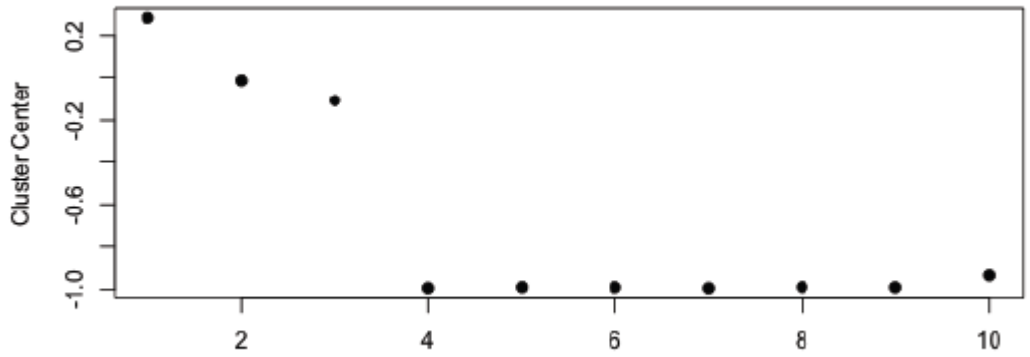
```
##
##      laying sitting standing walk walkdown walkup
##   1      29       0        0    0        0      0
##   2       3       0        0    0        0     53
##   3       0       0        0    0       49      0
##   4       0       0        0   95        0      0
##   5       0      37       51    0        0      0
##   6      18      10        2    0        0      0
```

Dobra separacja dla
Walk
Walkup
walkdown

Klaster 1: laying

- Gdzie leży centrum tego klastra w 500-dim przestrzeni, narysuj pierwsze 10 zmiennych

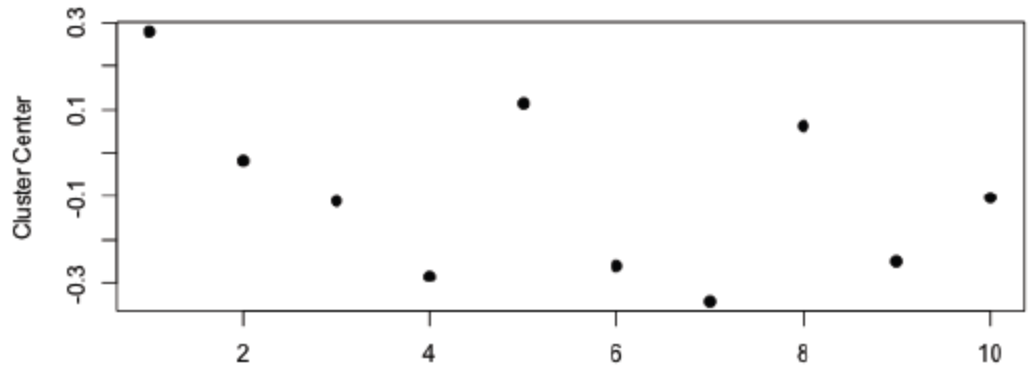
```
plot(kClust$center[1, 1:10], pch = 19, ylab = "Cluster Center", xlab = "")
```



Klaster 2: walking

- Gdzie leży centrum tego klastra w 500-dim przestrzeni, narysuj pierwsze 10 zmiennych

```
plot(kClust$center[4, 1:10], pch = 19, ylab = "Cluster Center", xlab = "")
```



Air-polution dane

- Czy teraz zanieczyszczenia są mniejsze niż dawniej?
  - Czy jest niższe teraz (2010) niż było w (1999)?
  - Czy fluktuacje się zmniejszyły?
- Porównaj dla kilku stanów
- Porównaj dla jednego stanu wybranego
- Porównaj dla całego USA

Uruchom i zrozum sekwencje w załączonym skrypcie.