Data Mining - Projekt PW 2013/2014

Marcin Kosiński, Marta Sommer,
kosinskim@student.mini.pw.edu.pl, sommerm@student.mini.pw.edu.pl

17 maja 2014

Spis treści

Spis rysunków

Spis tabel

1 Wstępna analiza danych

1.1 Motywacja

1.2 Opis danych

1.3 Podstawowe charakterystyki zmiennych

```
  seismic seismoacoustic shift genergy gpuls gdenergy gdpuls ghazard nbumps nbumps2 nbumps3 nbumps4
1       a              a     N   15180    48      -72    -72       a      0       0       0       0
2       a              a     N   14720    33      -70    -79       a      1       0       1       0
3       a              a     N    8050    30      -81    -78       a      0       0       0       0
4       a              a     N   28820   171      -23     40       a      1       0       1       0
5       a              a     N   12640    57      -63    -52       a      0       0       0       0
6       a              a     W   63760   195      -73    -65       a      0       0       0       0
  nbumps5 energy maxenergy class
1       0      0         0     0
2       0   2000      2000     0
3       0      0         0     0
4       0   3000      3000     0
5       0      0         0     0
6       0      0         0     0
```

1.3.1 Zmienne ilościowe

	genergy	gpuls	gdenergy	gdpuls	nbumps	nbumps2	nbumps3	nbumps4	nbumps5	energy	maxenergy
Odchylenie St	229200.51	562.65	80.32	63.17	1.36	0.78	0.77	0.28	0.07	20450.83	19357.45
Wariancja	52532873277.49	316577.88	6451.15	3990.01	1.86	0.61	0.59	0.08	0.00	418236579.49	374711059.50
Mediana	25485.00	379.00	-6.00	-6.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Średnia	90242.52	538.58	12.38	4.51	0.86	0.39	0.39	0.07	0.00	4975.27	4278.85

Tabela 1.1: Podstawowe statystyki agregujące dla zmiennych ciągłych.

```
summary(se[, -c(1:3, 8, 16)])

    genergy            gpuls          gdenergy          gdpuls          nbumps          nbumps2
 Min.   :    100   Min.   :   2   Min.   : -96.0   Min.   :-96.0   Min.   :0.00   Min.   :0.000
 1st Qu.:  11660   1st Qu.: 190   1st Qu.: -37.0   1st Qu.:-36.0   1st Qu.:0.00   1st Qu.:0.000
 Median :  25485   Median : 379   Median :  -6.0   Median : -6.0   Median :0.00   Median :0.000
 Mean   :  90243   Mean   : 539   Mean   :  12.4   Mean   :  4.5   Mean   :0.86   Mean   :0.394
 3rd Qu.:  52832   3rd Qu.: 669   3rd Qu.:  38.0   3rd Qu.: 30.2   3rd Qu.:1.00   3rd Qu.:1.000
 Max.   :2595650   Max.   :4518   Max.   :1245.0   Max.   :838.0   Max.   :9.00   Max.   :8.000
    nbumps3         nbumps4           nbumps5            energy         maxenergy
 Min.   :0.000   Min.   :0.0000   Min.   :0.0000   Min.   :     0   Min.   :     0
 1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:     0
 Median :0.000   Median :0.0000   Median :0.0000   Median :     0   Median :     0
 Mean   :0.393   Mean   :0.0677   Mean   :0.0046   Mean   :  4975   Mean   :  4279
 3rd Qu.:1.000   3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:  2600   3rd Qu.:  2000
 Max.   :7.000   Max.   :3.0000   Max.   :1.0000   Max.   :402000   Max.   :400000

apply(se[, -c(1:3, 8, 16)], 2, shapiro.test)

$genergy

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.3776, p-value < 2.2e-16


$gpuls

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.7571, p-value < 2.2e-16


$gdenergy

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.7828, p-value < 2.2e-16


$gdpuls

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.8462, p-value < 2.2e-16


$nbumps

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.6733, p-value < 2.2e-16


$nbumps2

Shapiro-Wilk normality test
```

```
data:  newX[, i]
W = 0.5626, p-value < 2.2e-16


$nbumps3

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.5674, p-value < 2.2e-16


$nbumps4

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.2522, p-value < 2.2e-16


$nbumps5

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.0397, p-value < 2.2e-16


$energy

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.2361, p-value < 2.2e-16


$maxenergy

Shapiro-Wilk normality test

data:  newX[, i]
W = 0.2058, p-value < 2.2e-16

cor(se[, -c(1:3, 8, 16)], method = "spearman")

          genergy   gpuls gdenergy  gdpuls  nbumps   nbumps2  nbumps3  nbumps4   nbumps5  energy
genergy   1.00000 0.76075 0.394387 0.39317 0.49179  0.344367 0.391806  0.22806  0.040279 0.47936
gpuls     0.76075 1.00000 0.459023 0.58075 0.30739  0.223082 0.208510  0.22138  0.041573 0.30422
gdenergy  0.39439 0.45902 1.000000 0.79950 0.07160  0.081121 0.006386  0.06719  0.055157 0.06797
gdpuls    0.39317 0.58075 0.799502 1.00000 0.09863  0.091861 0.036046  0.08879  0.066378 0.09705
nbumps    0.49179 0.30739 0.071599 0.09863 1.00000  0.759414 0.757964  0.34340  0.087215 0.95571
nbumps2   0.34437 0.22308 0.081121 0.09186 0.75941  1.000000 0.306138  0.14867 -0.002579 0.58583
nbumps3   0.39181 0.20851 0.006386 0.03605 0.75796  0.306138 1.000000  0.15657  0.058727 0.78851
nbumps4   0.22806 0.22138 0.067190 0.08879 0.34340  0.148671 0.156567  1.00000 -0.017429 0.45217
nbumps5   0.04028 0.04157 0.055157 0.06638 0.08722 -0.002579 0.058727 -0.01743  1.000000 0.13017
energy    0.47936 0.30422 0.067971 0.09705 0.95571  0.585830 0.788512  0.45217  0.130172 1.00000
maxenergy 0.46999 0.29729 0.064563 0.09326 0.94225  0.561806 0.783260  0.45435  0.130219 0.99766
          maxenergy
genergy     0.46999
gpuls       0.29729
gdenergy    0.06456
gdpuls      0.09326
nbumps      0.94225
```

```
nbumps2     0.56181
nbumps3     0.78326
nbumps4     0.45435
nbumps5     0.13022
energy      0.99766
maxenergy   1.00000
```

1.3.2 Zmienna objaśniana

```
summary(se[, 16])

   0    1
2414  170
```