Projekt Zaliczeniowy nr 2

Mateusz Kapusta

2022-05-30

Przygotowanie techniczne

Ze względu na dużą ilość obliczeń do wykonania skorzystamy z równoległego modelu wykonania obliczeń. W tym celu ustawiamy liczbę dostępnych wątków jako 90% dostępnych wątków oraz ustawiamy jako domyślną metodę fork-a ze standardu POSIX. Skorzystamy z biblioteki doParallel oraz foreach.

```
library(doParallel)
```

```
## Loading required package: foreach
```

```
## Loading required package: iterators
```

```
## Loading required package: parallel
```

```
n<-as.integer(parallel::detectCores()*0.9)
para<-parallel::makeCluster(n,type="FORK")
doParallel::registerDoParallel(cl = para)
foreach::getDoParRegistered()
```

```
## [1] TRUE
```

```
print(para)
```

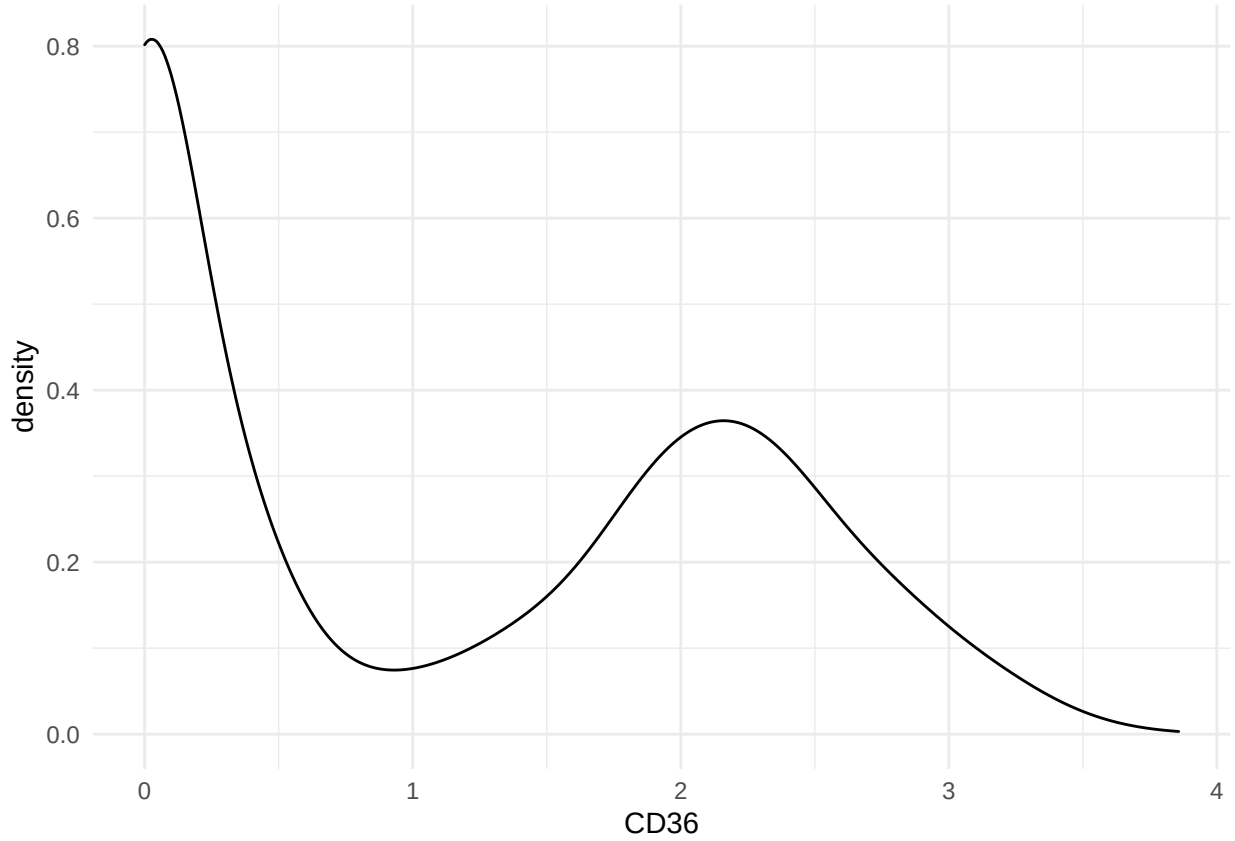
```
## socket cluster with 14 nodes on host 'localhost'
```

Eksploracja danych

```
xtrain<-read.csv("X_train.csv")
ytrain<-read.csv("y_train.csv")
xtest<-read.csv("X_test.csv")
```

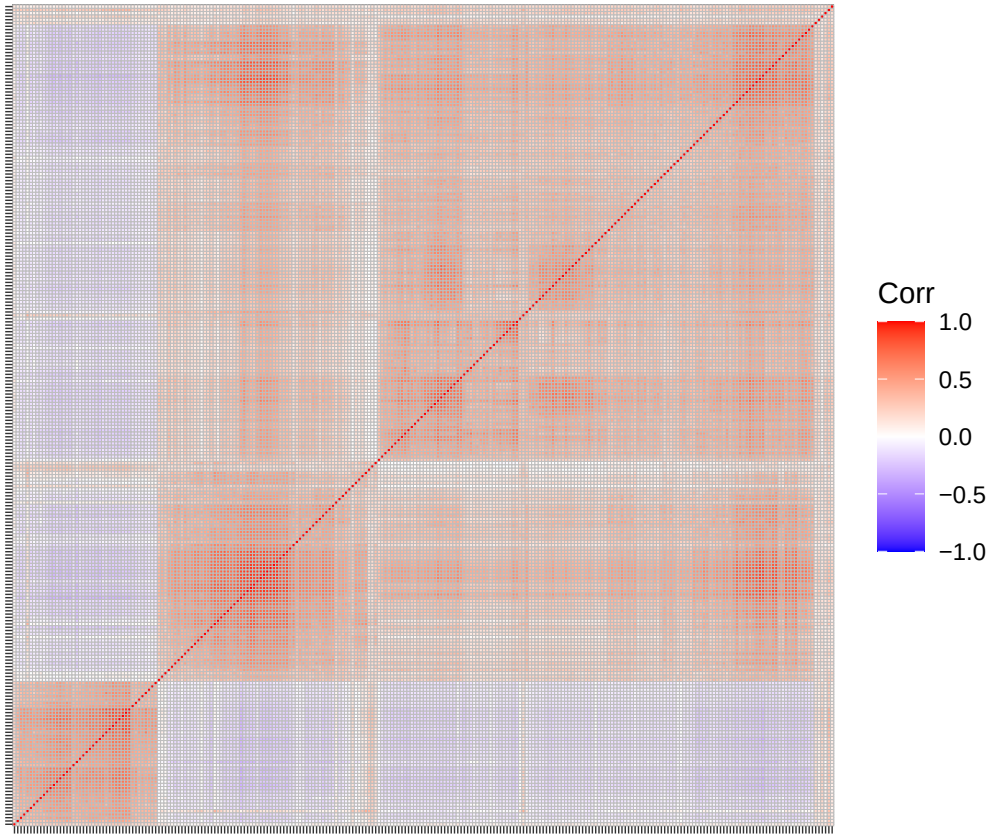
W naszych danych mamy 3794 obserwacji z czego każdej obserwacji odpowiada 9000 zmiennych objaśniających. W danych mamy 0 braków w danych. Zwizualizujmy teraz rozkład zmiennej objaśnianej.

```
ggplot(ytrain)+geom_density(aes(CD36))+theme_minimal()
```



Teraz zbadajmy najbardziej skorelowane zmienne ze zmienną objaśnianą.

```
core<-numeric(ncol(xtrain))
for (i in 1:ncol(xtrain))
{
  core[i]<-cor(ytrain$CD36,xtrain[,i])
}
indexy<-order(core, decreasing=TRUE)[1:250]
mat <- round(cor(xtrain[indexy]),3)
ggcorrplot(mat,"hc.order"=TRUE,
          ggtheme=ggplot2::theme_dark,
          tl.cex=0
          )
```



Elastic Net

Model elastic net charakteryzowany jest przez dwa parametry λ oraz α. Dla podanych argumentów celem jest minimalizacja

$$RSS + \lambda \left(\sum_i B_i^2 \frac{(1-\alpha)}{2} + \alpha \sum_i |\beta_i| \right) \tag{1}$$