lassopart

Nooreen S Dabbish

<2016-02-12 Fri>

Outline

Ensemble method based on lasso regression

- takes advantage of lasso's sparsity property
- index for scoring variable importance
  - scores based on subsampling and ensemble learning
  - using penalized lasso linear regression
- coordinate descent
  - GLMNET library
  - time-efficient use of full data space
- No feature reduction: full set of correlations

Matrix vectorization and adding responses (gender)

```
## Create list of unique region abbreviations from ori
## Adds integers (starting with 2) to non-unique value
setwd('~/Documents/BejingZhang')
abbrev <- read.table("abbreviations.txt")
abbrev[] <- lapply(abbrev, as.character)
v <- vector("list", 0)
v <- abbrev[[1]][1]
for (j in 2:dim(abbrev)[1]){
    x <- 1
    y <- abbrev[[1]][j]
    while (is.element(y, v)){
       x <- x+1
       y <- paste0(abbrev[[1]][j],as.character(x), sep
    }
    v <- c(v,y)
}
```

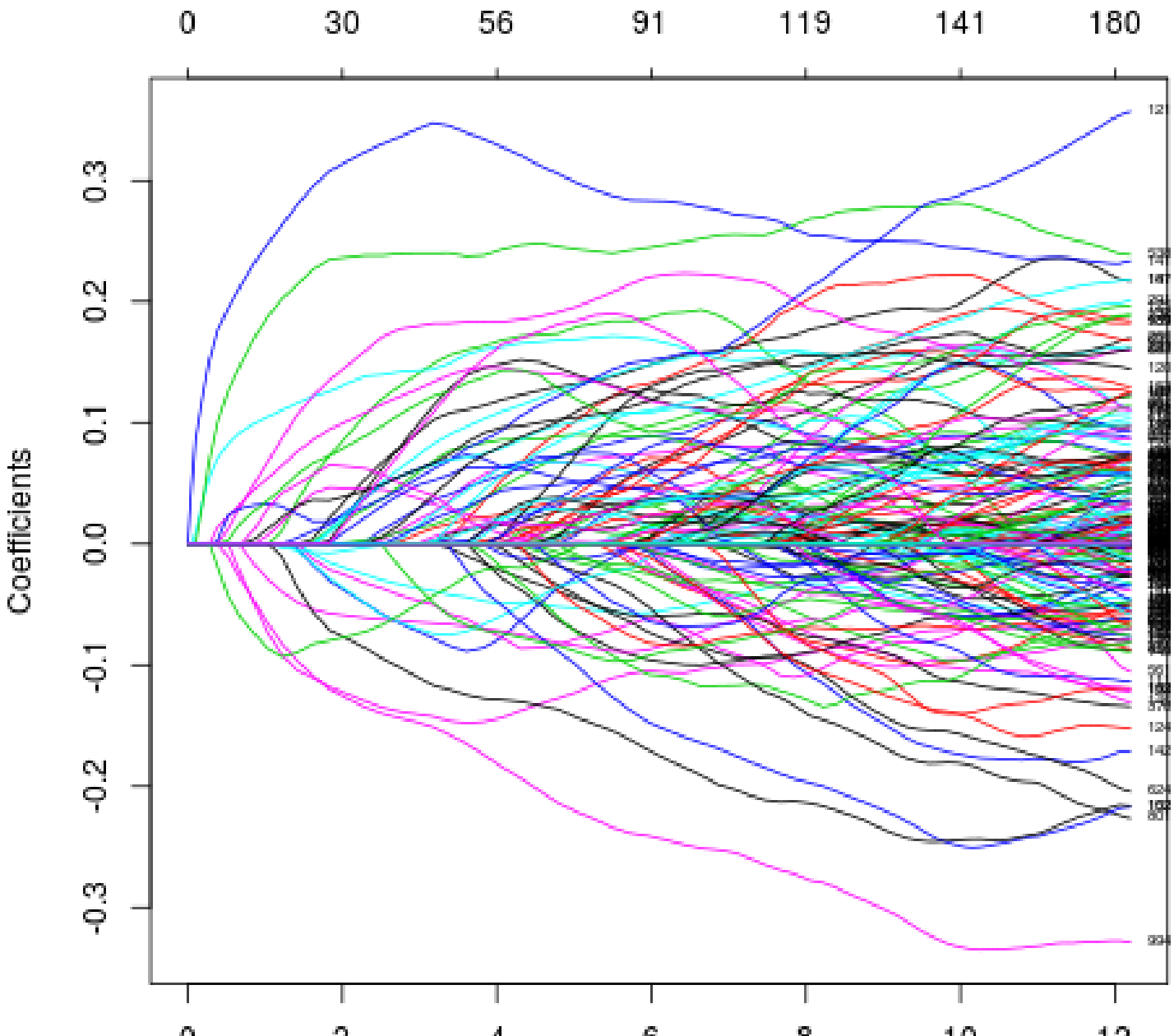
Code

```
## install.packages("glmnet", repos='http://cran.us.r-p
library(glmnet)

fit = glmnet(x = as.matrix(vectorgender[,-(1:3)]),
             y = as.vector(vectorgender$gender))

plot(fit, label =TRUE)
```

Plot

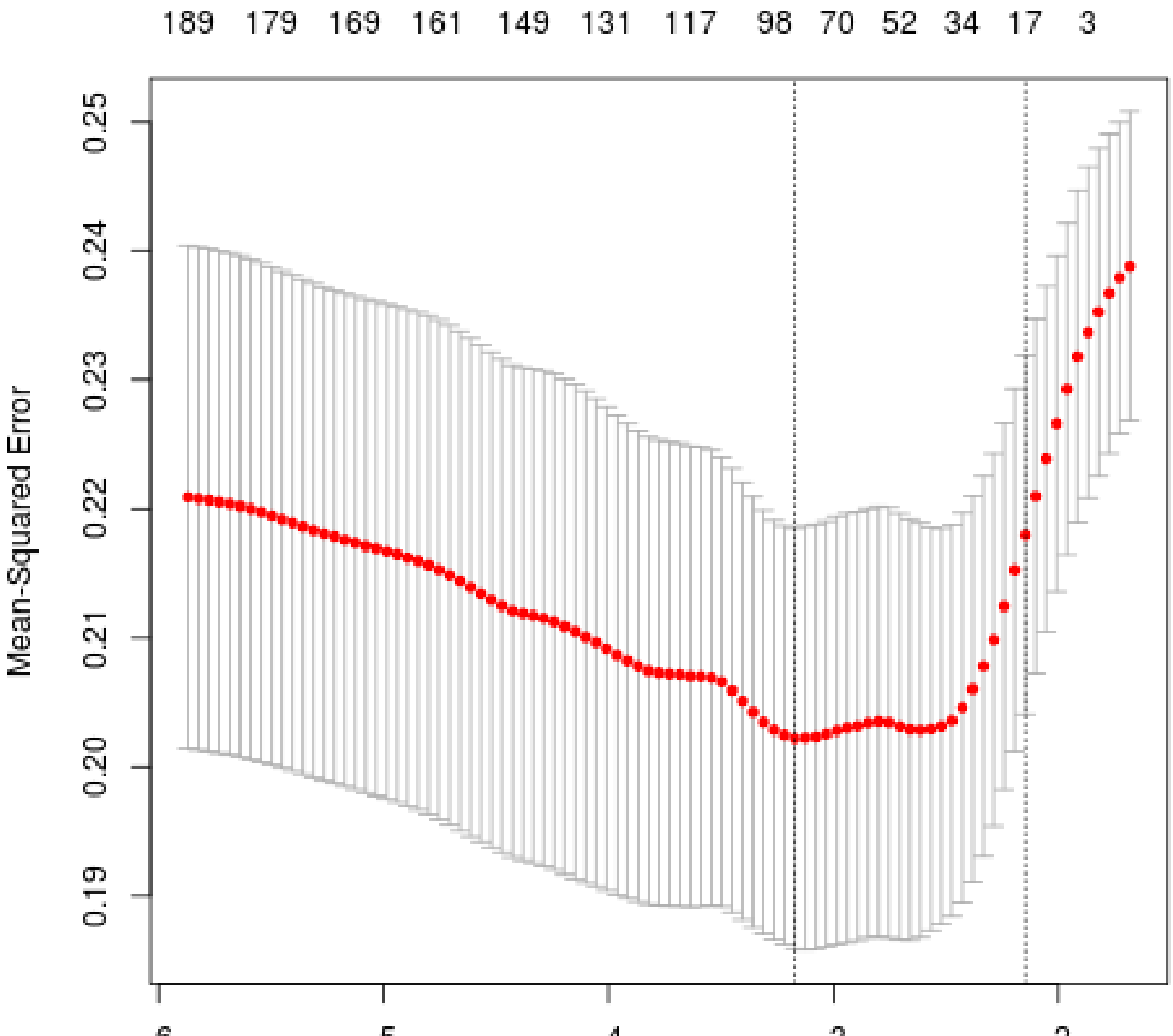


Code

```
cvfit = cv.glmnet(x = as.matrix(vectorgender[,-(1:3)]),
                  y = (vectorgender$gender))

plot(cvfit)
```

Plot


189 179 169 161 149 131 117 98 70 52 34 17 3
Mean-Squared Error
0.19
0.20
0.21
0.22
0.23
0.24
0.25

Selected λ s

```
cvfit$lambda.min
```

```
[1] 0.07332683
```

`lambda.min` value that gives minimum cross-validated error

```
cvfit$lambda.1se
```

```
[1] 0.09252797
```

`lambda.1se` value that gives most regularized model with error vwithin one standard error of minimum cross-validated error