Homework

Martyna Plomecka

9/30/2020

Big Data Science: Assignment 1

This assignment is based on the R tutorial code on regression, crossvalidation, and regularization. The data used in the tutorial are Big Five personality trait questionnaire scores from 2800 students, including thei age, education, and gender. There are a total of 5 assignments. Grading is based on the correctness of the # output variables and the code used to produce the output.

```
#load required packages

if(!require("psych")){
  install.packages("psych")
  library("psych")
}
```

```
## Loading required package: psych
```

```
if(!require("glmnet")){
  install.packages("glmnet")
  library("glmnet")
}
```

```
## Loading required package: glmnet
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.0-2
```

```
if(!require("Matrix")){
  install.packages("Matrix")
  library("Matrix")
}
```

load the dataset from: https://www.personality-project.org/r/html/bfi.html

```
data('bfi')
```

Data preparation:

```
#remove participants with missing values
ind = rowSums(is.na(bfi))==0
bfi = bfi[ind,]

#Features: Big Five questionnaire scores
X = bfi[,c(1:5,6:10,11:15,16:20,21:25)]
rownames(X)= NULL
X = data.matrix(X, rownames.force = NA)
```

```r
#Labels: Age
y = bfi[,c(28)]
rownames(y)= NULL
y = data.matrix(y, rownames.force = NA)
y = rowMeans(y)
```

Task 1

Inspect your data. Return the following variables "nfeatures" (number of features), "nsamples" (number of samples), "minlabel" (minimum value across labels), "maxlabel" (maximum value across labels).

```r
# Your Response Here
```

Task 2

Predict y from X using a linear regression (no regularization, no crossvalidation). Return a variable "rsquared" that contains the R-squared value of the model fit with at least four-digit precision (e.g. rsquared = 0.9752).

```r
# Your Response Here
```

Task 3

Inspect the coefficients of the fitted LASSO model (e.g. "fit$beta") and return a variable "ncoeff" that contains the number of non-zero coefficients in the model.

```r
# Your Response Here
```

Task 4

Predict y from X using a RIDGE regression (no crossvalidation). Hint: The difference between RIDGE and LASSO regresssion in glmntet is in the choice of the alpha parameter (Use the internet to learn more about the alpha parameter in glmnet, and about the difference between RIDGE and LASSO). Your task is to compare the coefficients two different RIDGE model model fits. The first model uses less regularization (lambda=0.01) and the second model more regularization (lambda=1). Return two variables "less_reg" and "more_reg" that each contain the minimum, maximum, and max-min difference of coefficient values for each case, using at least a four-digit precision. Example: coefficients=c(-2,0,-5,4,1), return_variable=c(-5.0000,4.0000,9.0000).

```r
# Your Response Here
```

Task 5

Predict y from X using a LASSO regression with crossvalidation (glmnet default parameters for lambda and cv). What is the maximum number of coefficients that can be dropped (set to zero) before LASSO looses fit performance over linear regression? The loss of performance is defined as LASSO Mean-Squared Error confidence intervals exceeding the Mean-Squared Error of the linear regression. Return a variable "ndrop" that contains the number of coefficients that can be dropped.

```r
# Your Response Here
```