HW_DAY20_Andhyka Cakrabuana Adhitama

Dhyka

12/12/2021

Preparation

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(caTools)
library(psych)
library(glmnet)
```

```
## Loading required package: Matrix

## Loaded glmnet 4.1-3
```

```
dfori <- read.csv("https://raw.githubusercontent.com/pararawendy/dibimbing-materials/main/boston.csv")
head(dfori, 5)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33
##   medv
## 1 24.0
## 2 21.6
## 3 34.7
## 4 33.4
## 5 36.2
```

1. Split data: train - validate - test

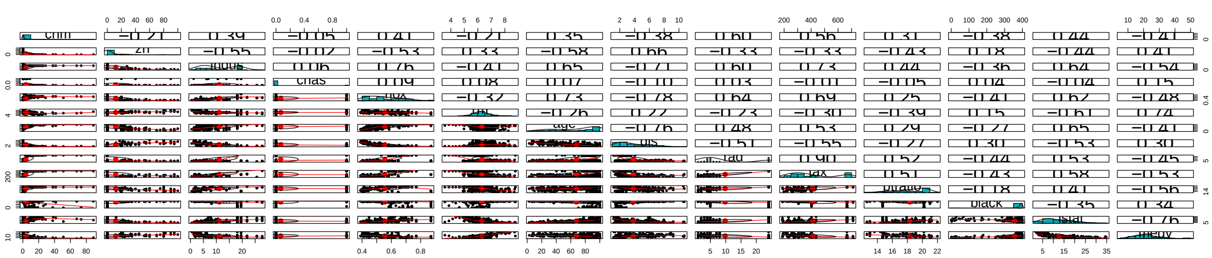
```
set.seed(123)
sample <- sample.split(dfori$medv, SplitRatio = .80)
pre_train <- subset(dfori, sample == TRUE)
sample_train <- sample.split(pre_train$medv, SplitRatio = .80)
```

Train-Validation-Test

```
train <- subset(pre_train, sample_train == TRUE)
validation <- subset(pre_train, sample_train == FALSE)
test <-subset(dfori, sample == FALSE)
```

2. Draw correlation plot on training data and perform feature selection on highly correlated features

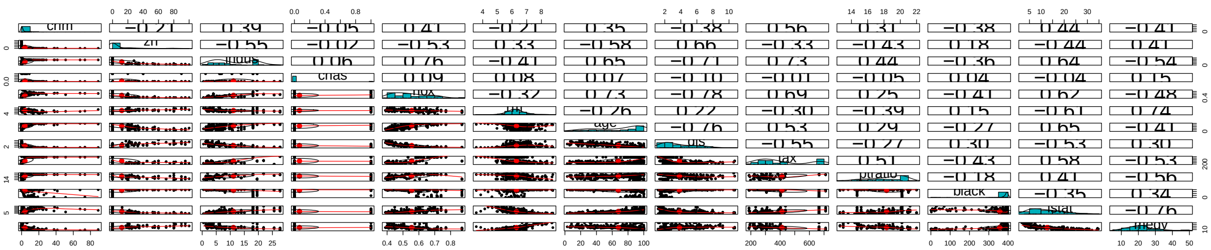
```
pairs.panels(train,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
             density = TRUE,  # show density plots
             ellipses = TRUE ) # show correlation ellipses
```



Drop correlated columns

```
# From the plot above, we know that :
# rad & tax are highly correlated each other (0,9), so i decide to go with tax (-0.53) and drop rad (-0
# Note : treshold: absolute(corr)>0.8
train <- train %>% select(-'rad')
validation <- validation %>% select(-'rad')
test <- test %>% select(-'rad')
pairs.panels(train,
             method = "pearson", # correlation method
             hist.col = "#00AFBB",
```

```
              density = TRUE,  # show density plots
              ellipses = TRUE ) # show correlation ellipses
```



3. Fit models on training data (lambdas = [0.01, 0.1, 1, 10])

Feature preprocessing

```
x <- model.matrix(medv ~ ., train)[,-1]
y <- train$medv
```

Ridge Regression

```
#lambda 0.01
ridge_reg_pointzeroone <- glmnet(x, y, alpha = 0, lambda = 0.01)
coef(ridge_reg_pointzeroone)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept)  2.807966e+01
## crim        -7.972347e-02
## zn           3.796482e-02
## indus       -4.106178e-02
## chas         2.893259e+00
## nox         -1.602703e+01
## rm           4.517287e+00
## age          5.679736e-03
## dis         -1.314253e+00
## tax         -2.421124e-04
## ptratio     -9.031044e-01
## black        6.572154e-03
## lstat       -4.779743e-01
```

```
#lambda 0.1
ridge_reg_pointone <- glmnet(x, y, alpha = 0, lambda = 0.1)
coef(ridge_reg_pointone)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  2.720583e+01
## crim        -7.865999e-02
## zn           3.682165e-02
## indus       -4.208117e-02
## chas         2.888898e+00
## nox         -1.513326e+01
## rm           4.524625e+00
## age          5.018603e-03
## dis         -1.260076e+00
## tax         -4.973179e-04
## ptratio     -8.931536e-01
## black        6.639672e-03
## lstat       -4.709285e-01
```

```r
#lambda 1
ridge_reg_one <- glmnet(x, y, alpha = 0, lambda = 1)
coef(ridge_reg_one)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                      s0
## (Intercept) 22.53185712
## crim        -0.07288298
## zn           0.02967177
## indus       -0.05282228
## chas         2.84365458
## nox         -9.91885353
## rm           4.44323173
## age          0.00101080
## dis         -0.90469612
## tax         -0.00195283
## ptratio     -0.82592673
## black        0.00688591
## lstat       -0.41785676
```

```r
#lambda 10
ridge_reg_ten <- glmnet(x, y, alpha = 0, lambda = 10)
coef(ridge_reg_ten)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept) 21.798954932
## crim        -0.061446034
## zn           0.020349479
## indus       -0.081406215
## chas         2.105932529
## nox         -4.343751697
## rm           2.889836712
## age         -0.008076179
## dis         -0.190031642
## tax         -0.003586798
## ptratio     -0.571970624
```

```
## black        0.005615501
## lstat       -0.242577448
```

Lasso Regression

```r
#lambda 0.01
lasso_reg_pointzeroone <- glmnet(x, y, alpha = 1, lambda = 0.01)
coef(lasso_reg_pointzeroone)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  2.782960e+01
## crim        -7.879101e-02
## zn           3.673332e-02
## indus       -3.849552e-02
## chas         2.864983e+00
## nox         -1.574647e+01
## rm           4.531757e+00
## age          4.411184e-03
## dis         -1.294476e+00
## tax         -2.439776e-04
## ptratio     -9.039250e-01
## black        6.556538e-03
## lstat       -4.764532e-01
```

```r
#lambda 0.1
lasso_reg_pointone <- glmnet(x, y, alpha = 1, lambda = 0.1)
coef(lasso_reg_pointone)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  2.472929e+01
## crim        -6.891240e-02
## zn           2.563768e-02
## indus       -1.728300e-02
## chas         2.590973e+00
## nox         -1.267687e+01
## rm           4.620427e+00
## age          .
## dis         -1.022117e+00
## tax         -5.088209e-04
## ptratio     -9.019518e-01
## black        6.368681e-03
## lstat       -4.677220e-01
```

```r
#lambda 1
lasso_reg_one <- glmnet(x, y, alpha = 1, lambda = 1)
coef(lasso_reg_one)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
```

```
##                        s0
## (Intercept) 13.987998322
## crim        -0.010183404
## zn           .
## indus        .
## chas         .
## nox          .
## rm           4.306536549
## age          .
## dis          .
## tax         -0.000775465
## ptratio     -0.710899295
## black        0.001632894
## lstat       -0.457861803
```

```
#lambda 10
lasso_reg_ten <- glmnet(x, y, alpha = 1, lambda = 10)
coef(lasso_reg_ten)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                    s0
## (Intercept) 22.53775
## crim         0.00000
## zn           .
## indus        .
## chas         .
## nox          .
## rm           .
## age          .
## dis          .
## tax          .
## ptratio      .
## black        .
## lstat        .
```

4. Choose best lambda from validation set

Features

```
x_validation <- model.matrix(medv ~., validation)[,-1]
y_validation <- validation$medv
```

Ridge regression (using RMSE)

```
RMSE_ridge_pointzeroone <- sqrt(mean((y_validation - predict(ridge_reg_pointzeroone, x_validation))^2))
RMSE_ridge_pointzeroone # 4.3464 -> best
```

```
## [1] 4.3464
```

```r
RMSE_ridge_pointone <- sqrt(mean((y_validation - predict(ridge_reg_pointone, x_validation))^2))
RMSE_ridge_pointone # 4.3495
```

```
## [1] 4.349494
```

```r
RMSE_ridge_one <- sqrt(mean((y_validation - predict(ridge_reg_one, x_validation))^2))
RMSE_ridge_one # 4.4220
```

```
## [1] 4.422032
```

```r
RMSE_ridge_ten <- sqrt(mean((y_validation - predict(ridge_reg_ten, x_validation))^2))
RMSE_ridge_ten # 5.3421
```

```
## [1] 5.342122
```

```r
# callback best lambda (0.01)
ridge_reg_pointzeroone <- glmnet(x, y, alpha = 0, lambda = 0.01)
coef(ridge_reg_pointzeroone)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  2.807966e+01
## crim        -7.972347e-02
## zn           3.796482e-02
## indus       -4.106178e-02
## chas         2.893259e+00
## nox         -1.602703e+01
## rm           4.517287e+00
## age          5.679736e-03
## dis         -1.314253e+00
## tax         -2.421124e-04
## ptratio     -9.031044e-01
## black        6.572154e-03
## lstat       -4.779743e-01
```

Medv = 20.808 - 0.080 crim + 0.038 zn - 0.041 indus + 2.893 chas - 16.027 nox + 4.517 rm + 0.006 age - 1.314 dis - 0.0002 tax - 0.903 ptratio + 0.007 black - 0.477 lstat

```r
# Interpretation
# With fixed point 20.808 in medv,
# An increase of 1 point in crim, while the other features are kept fixed, is associated with an decrea
# An increase of 1 point in zn, while other features are kept fixed, is associated with an increase of
# An increase of 1 point in indus, while other features are kept fixed, is associated with an decrease
# An increase of 1 point in chas, while other features are kept fixed, is associated with an increase o
# An increase of 1 point in nox, while other features are kept fixed, is associated with an decrease of
```

```
# An increase of 1 point in rm, while other features are kept fixed, is associated with an increase of
# An increase of 1 point in age, while other features are kept fixed, is associated with an increase of
# An increase of 1 point in dis, while other features are kept fixed, is associated with an decrease of
# An increase of 1 point in tax, while other features are kept fixed, is associated with an decrease of
# An increase of 1 point in ptratio, while other features are kept fixed, is associated with an decreas
# An increase of 1 point in black, while other features are kept fixed, is associated with an increase
# An increase of 1 point in lstat, while other features are kept fixed, is associated with an  decrease
```

Lasso regressiong (using RMSE)

```
RMSE_lasso_pointzeroone <- sqrt(mean((y_validation - predict(lasso_reg_pointzeroone, x_validation))^2))
RMSE_lasso_pointzeroone # 4.3408 -> best
```

```
## [1] 4.340783
```

```
RMSE_lasso_pointone <- sqrt(mean((y_validation - predict(lasso_reg_pointone, x_validation))^2))
RMSE_lasso_pointone # 4.3527
```

```
## [1] 4.352728
```

```
RMSE_lasso_one <- sqrt(mean((y_validation - predict(lasso_reg_one, x_validation))^2))
RMSE_lasso_one # 4.9378
```

```
## [1] 4.937774
```

```
RMSE_lasso_ten <- sqrt(mean((y_validation - predict(lasso_reg_ten, x_validation))^2))
RMSE_lasso_ten # 9.371755
```

```
## [1] 9.371755
```

```
# callback best lambda (0.01)
lasso_reg_pointone <- glmnet(x, y, alpha = 1, lambda = 0.1)
coef(lasso_reg_pointzeroone)
```

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)  2.782960e+01
## crim        -7.879101e-02
## zn           3.673332e-02
## indus       -3.849552e-02
## chas         2.864983e+00
## nox         -1.574647e+01
## rm           4.531757e+00
## age          4.411184e-03
## dis         -1.294476e+00
## tax         -2.439776e-04
## ptratio     -9.039250e-01
## black        6.556538e-03
## lstat       -4.764532e-01
```

Medv = 20.783 - 0.079 crim + 0.037 zn - 0.038 indus + 2.865 chas - 15.746 nox + 4.532 rm + 0.004 age - 1.294 dis - 0.0002 tax - 0.904 ptratio + 0.007 black - 0.476 lstat

```
# Interpretation
# With fixed point 20.783 in medv,
# An increase of 1 point in crim, while the other features are kept fixed, is associated with an decrea
# An increase of 1 point in zn, while other features are kept fixed, is associated with an increase of
# An increase of 1 point in indus, while other features are kept fixed, is associated with an decrease
# An increase of 1 point in chas, while other features are kept fixed, is associated with an increase o
# An increase of 1 point in nox, while other features are kept fixed, is associated with an decrease of
# An increase of 1 point in rm, while other features are kept fixed, is associated with an increase of
# An increase of 1 point in age, while other features are kept fixed, is associated with an increase of
# An increase of 1 point in dis, while other features are kept fixed, is associated with an decrease of
# An increase of 1 point in tax, while other features are kept fixed, is associated with an decrease of
# An increase of 1 point in ptratio, while other features are kept fixed, is associated with an decreas
# An increase of 1 point in black, while other features are kept fixed, is associated with an increase
# An increase of 1 point in lstat, while other features are kept fixed, is associated with an  decrease
```

5. Evaluate the best models on the test data (+interpretation)

Feature

```
x_test <- model.matrix(medv ~., test)[,-1]
y_test <- test$medv
```

Ridge

```
RMSE_ridge_best <- sqrt(mean((y_test - predict(ridge_reg_pointzeroone, x_test))^2))
RMSE_ridge_best
```

```
## [1] 6.820639
```

The standard deviation of prediction errors is 6.820 i.e. from the regression line, the residuals mostly deviate between +- 6.820

```
MAE_ridge_best <- mean(abs(y_test-predict(ridge_reg_pointzeroone, x_test)))
MAE_ridge_best
```

```
## [1] 3.896186
```

On average, the prediction deviates the true medv by 3.896

```
MAPE_ridge_best <- mean(abs((predict(ridge_reg_pointzeroone, x_test) - y_test))/y_test)
MAPE_ridge_best
```

```
## [1] 0.1710101
```

Moreover, this 3.896 is equivalent to 17.10% deviation relative to the true medv

Lasso

```
RMSE_lasso_best <- sqrt(mean((y_test - predict(lasso_reg_pointone, x_test))^2))
RMSE_lasso_best
```

```
## [1] 6.84918
```

The standard deviation of prediction errors is 6.849 i.e. from the regression line, the residuals mostly deviate between +- 6.849

```
MAE_lasso_best <- mean(abs(y_test-predict(lasso_reg_pointone, x_test)))
MAE_lasso_best
```

```
## [1] 3.819946
```

On average, the prediction deviates the true medv by 3.820

```
MAPE_lasso_best <- mean(abs((predict(lasso_reg_pointone, x_test) - y_test))/y_test)
MAPE_lasso_best
```

```
## [1] 0.1680723
```

Moreover, this 3.819 is equivalent to 16.80% deviation relative to the true medv