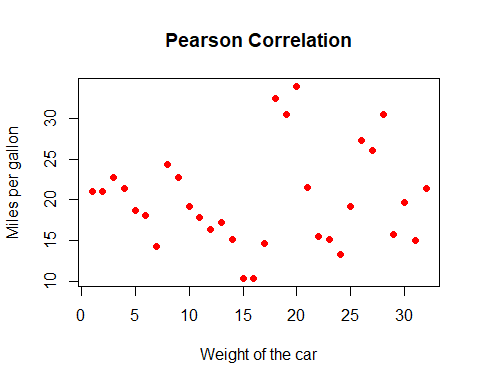
Feature Engineering

Turner Sale

11/9/2019

Plotting weight vs. MPG

plot(mtcars$mpg, col = 'red', xlab = 'Weight of the car', ylab = 'Miles per gallon', pch = 16, main = 'Pearson Correlation')



Correlate these two values

cor(mtcars, method = 'pearson')

## mpg cyl disp hp drat wt  
## mpg 1.0000000 -0.8521620 -0.8475514 -0.7761684 0.68117191 -0.8676594  
## cyl -0.8521620 1.0000000 0.9020329 0.8324475 -0.69993811 0.7824958  
## disp -0.8475514 0.9020329 1.0000000 0.7909486 -0.71021393 0.8879799  
## hp -0.7761684 0.8324475 0.7909486 1.0000000 -0.44875912 0.6587479  
## drat 0.6811719 -0.6999381 -0.7102139 -0.4487591 1.00000000 -0.7124406  
## wt -0.8676594 0.7824958 0.8879799 0.6587479 -0.71244065 1.0000000  
## qsec 0.4186840 -0.5912421 -0.4336979 -0.7082234 0.09120476 -0.1747159  
## vs 0.6640389 -0.8108118 -0.7104159 -0.7230967 0.44027846 -0.5549157  
## am 0.5998324 -0.5226070 -0.5912270 -0.2432043 0.71271113 -0.6924953  
## gear 0.4802848 -0.4926866 -0.5555692 -0.1257043 0.69961013 -0.5832870  
## carb -0.5509251 0.5269883 0.3949769 0.7498125 -0.09078980 0.4276059  
## qsec vs am gear carb  
## mpg 0.41868403 0.6640389 0.59983243 0.4802848 -0.55092507  
## cyl -0.59124207 -0.8108118 -0.52260705 -0.4926866 0.52698829  
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692 0.39497686  
## hp -0.70822339 -0.7230967 -0.24320426 -0.1257043 0.74981247  
## drat 0.09120476 0.4402785 0.71271113 0.6996101 -0.09078980  
## wt -0.17471588 -0.5549157 -0.69249526 -0.5832870 0.42760594  
## qsec 1.00000000 0.7445354 -0.22986086 -0.2126822 -0.65624923  
## vs 0.74453544 1.0000000 0.16834512 0.2060233 -0.56960714  
## am -0.22986086 0.1683451 1.00000000 0.7940588 0.05753435  
## gear -0.21268223 0.2060233 0.79405876 1.0000000 0.27407284  
## carb -0.65624923 -0.5696071 0.05753435 0.2740728 1.00000000

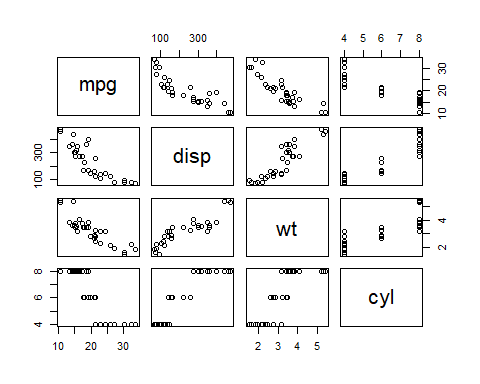
Correlation test

cor.test(x = mtcars$wt, y = mtcars$mpg)

##   
## Pearson's product-moment correlation  
##   
## data: mtcars$wt and mtcars$mpg  
## t = -9.559, df = 30, p-value = 1.294e-10  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.9338264 -0.7440872  
## sample estimates:  
## cor   
## -0.8676594

Plot variable correlations

pairs(~mpg + disp + wt + cyl, data=mtcars)



Iris dataset mutual information calculation

library("FSelector")

## Error: package or namespace load failed for 'FSelector':  
## .onLoad failed in loadNamespace() for 'rJava', details:  
## call: fun(libname, pkgname)  
## error: JAVA\_HOME cannot be determined from the Registry

ig\_values <- information.gain(Species~., iris)

## Error in information.gain(Species ~ ., iris): could not find function "information.gain"

ig\_values

## Error in eval(expr, envir, enclos): object 'ig\_values' not found

The error from above is due to incompatible Java versions (require the 64 bit version for other uses and cannot uninstall)

Wrapper method

library(caret)

## Loading required package: lattice

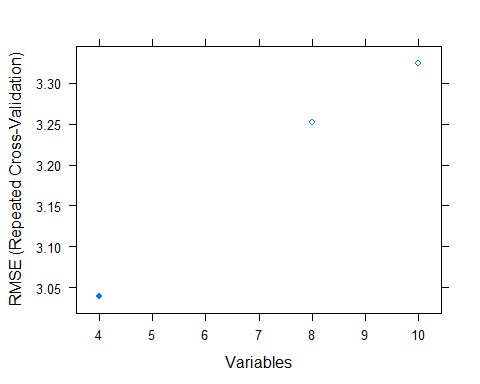
## Loading required package: ggplot2

rfe\_controller <- rfeControl(functions=lmFuncs, method="repeatedcv", repeats = 5, verbose = FALSE)  
size <- c(1:10)   
lm\_Profiler <- rfe(mtcars[,2:11], mtcars[,1], rfeControl = rfe\_controller)  
lm\_Profiler

##   
## Recursive feature selection  
##   
## Outer resampling method: Cross-Validated (10 fold, repeated 5 times)   
##   
## Resampling performance over subset size:  
##   
## Variables RMSE Rsquared MAE RMSESD RsquaredSD MAESD Selected  
## 4 3.039 0.8863 2.628 1.169 0.1706 1.082 \*  
## 8 3.252 0.8490 2.729 1.267 0.2502 1.040   
## 10 3.325 0.8474 2.854 1.284 0.2570 1.093   
##   
## The top 4 variables (out of 4):  
## wt, am, drat, gear

Plot profiler

plot(lm\_Profiler)



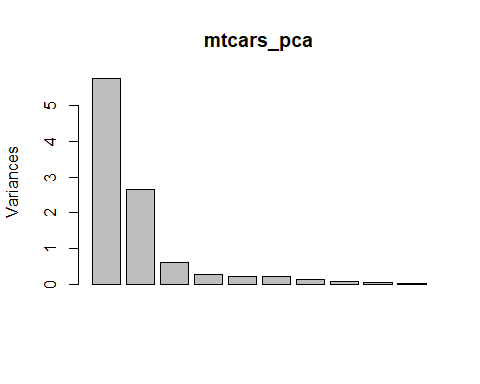
PCA

mtcars\_feat <- mtcars[,2:11]  
mtcars\_mpg <- mtcars[,1]  
mtcars\_pca <- prcomp(x = mtcars\_feat, scale. = T)  
mtcars\_pca

## Standard deviations (1, .., p=10):  
## [1] 2.4000453 1.6277725 0.7727968 0.5191403 0.4714341 0.4583857 0.3645821  
## [8] 0.2840450 0.2316298 0.1542606  
##   
## Rotation (n x k) = (10 x 10):  
## PC1 PC2 PC3 PC4 PC5  
## cyl 0.4029711 -0.03901479 0.13874360 -8.040022e-05 0.06148048  
## disp 0.3959243 0.05393117 0.01633491 -2.646304e-01 0.33851109  
## hp 0.3543255 -0.24496137 -0.18225874 6.000387e-02 0.52828704  
## drat -0.3155948 -0.27847781 -0.13057734 -8.528509e-01 0.10299748  
## wt 0.3668004 0.14675805 -0.38579961 -2.527210e-01 -0.14410292  
## qsec -0.2198982 0.46066271 -0.40307004 -7.174202e-02 -0.21341845  
## vs -0.3333571 0.22751987 -0.41252247 2.119502e-01 0.62369179  
## am -0.2474991 -0.43201042 0.23493804 3.190779e-02 0.04930286  
## gear -0.2214375 -0.46516217 -0.27929375 2.623809e-01 0.02039816  
## carb 0.2267080 -0.41169300 -0.56172255 1.233534e-01 -0.36576403  
## PC6 PC7 PC8 PC9 PC10  
## cyl -0.18206407 -0.04257067 0.07041306 -0.863268748 0.1670687388  
## disp 0.35738419 0.19767431 -0.14361684 -0.020039738 -0.6838300858  
## hp -0.03269674 -0.08503414 0.58708325 0.291428365 0.2462606844  
## drat -0.23386885 0.03226657 0.04010725 -0.086765162 0.0544414772  
## wt 0.43201764 -0.03368560 -0.36605124 0.075971836 0.5318885631  
## qsec 0.29265169 -0.03797611 0.59621869 -0.244573292 -0.1545795278  
## vs -0.11710663 -0.23387904 -0.36246041 -0.182200371 -0.0055443849  
## am 0.60874338 -0.54631997 0.02588771 -0.154149509 -0.0003995261  
## gear 0.24560902 0.69429321 -0.01069942 -0.198369367 0.0741152014  
## carb -0.25782743 -0.33623769 -0.08067483 0.003086198 -0.3585136181

Plot PCA

plot(mtcars\_pca)



Iris LDA

library('MASS')  
head(iris)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 setosa  
## 2 4.9 3.0 1.4 0.2 setosa  
## 3 4.7 3.2 1.3 0.2 setosa  
## 4 4.6 3.1 1.5 0.2 setosa  
## 5 5.0 3.6 1.4 0.2 setosa  
## 6 5.4 3.9 1.7 0.4 setosa

train <- sample(1:150, 75)  
lda\_model <- lda(Species ~., data = iris, subset=train)  
lda\_model$means

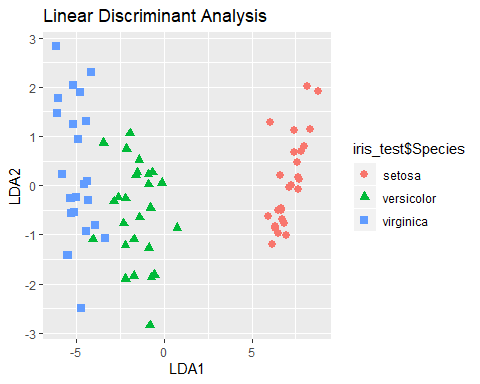
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## setosa 5.091667 3.445833 1.475000 0.2708333  
## versicolor 5.925000 2.675000 4.200000 1.2791667  
## virginica 6.670370 3.000000 5.622222 2.0740741

Predictions of LDA

lda\_pred <- predict(object = lda\_model, newdata = iris[-train, ])  
lda\_pred$class

## [1] setosa setosa setosa setosa setosa setosa   
## [7] setosa setosa setosa setosa setosa setosa   
## [13] setosa setosa setosa setosa setosa setosa   
## [19] setosa setosa setosa setosa setosa setosa   
## [25] setosa setosa versicolor versicolor versicolor versicolor  
## [31] versicolor versicolor versicolor versicolor versicolor versicolor  
## [37] versicolor virginica versicolor versicolor versicolor versicolor  
## [43] virginica versicolor versicolor versicolor versicolor versicolor  
## [49] versicolor versicolor versicolor versicolor virginica virginica   
## [55] virginica virginica virginica virginica virginica virginica   
## [61] virginica virginica virginica virginica virginica versicolor  
## [67] virginica virginica virginica virginica virginica virginica   
## [73] virginica virginica virginica   
## Levels: setosa versicolor virginica

iris\_test <- iris[-train, ]  
ggplot() + geom\_point(aes(lda\_pred$x[,1], lda\_pred$x[,2], colour = iris\_test$Species, shape = iris\_test$Species),   
size = 2.5) + ggtitle("Linear Discriminant Analysis") + xlab("LDA1") + ylab("LDA2") + labs(fill = "Species")



Embedded methods

library('randomForest')

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

form <- ~ wt + mpg + cyl  
rf\_model <- randomForest(formula = form, data = mtcars, importance = TRUE )  
rf\_model$importance

## 1 2 MeanDecreaseAccuracy MeanDecreaseGini  
## wt 0.1345141 -0.005187781 0.06345732 8.954729  
## mpg 0.1673415 -0.018039375 0.07393412 9.277803  
## cyl 0.1879763 0.007053567 0.09556791 3.842398

varImpPlot(rf\_model)

