Data624 - Homework7

Amit Kapoor

3/28/2021

Contents


```
library(AppliedPredictiveModeling)
library(tidyverse)
library(caret)
```

Exercise 6.2

Developing a model to predict permeability (see Sect. 1.4) could save significant resources for a pharmaceutical company, while at the same time more rapidly identifying molecules that have a sufficient permeability to become a drug:

(a)

Start R and use these commands to load the data.

The matrix fingerprints contains the 1,107 binary molecular predictors for the 165 compounds, while `permeability` contains permeability response.

```
data(permeability)
```

```
glimpse(permeability)
```

```
##  num [1:165, 1] 12.52 1.12 19.41 1.73 1.68 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:165] "1" "2" "3" "4" ...
##   ..$ : chr "permeability"
```

```
nrow(permeability)
```

```
## [1] 165
```

```
glimpse(fingerprints)
```

```
##  num [1:165, 1:1107] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:165] "1" "2" "3" "4" ...
##   ..$ : chr [1:1107] "X1" "X2" "X3" "X4" ...
```

```
ncol(fingerprints)
```

```
## [1] 1107
```

```
nrow(fingerprints)
```

```
## [1] 165
```

In this data there were 165 unique compounds; 1107 molecular fingerprints were determined for each. A molecular fingerprint is a binary sequence of numbers that represents the presence or absence of a specific molecular sub-structure. The response is highly skewed, the predictors are sparse (15.5 percent are present), and many predictors are strongly associated

(b)

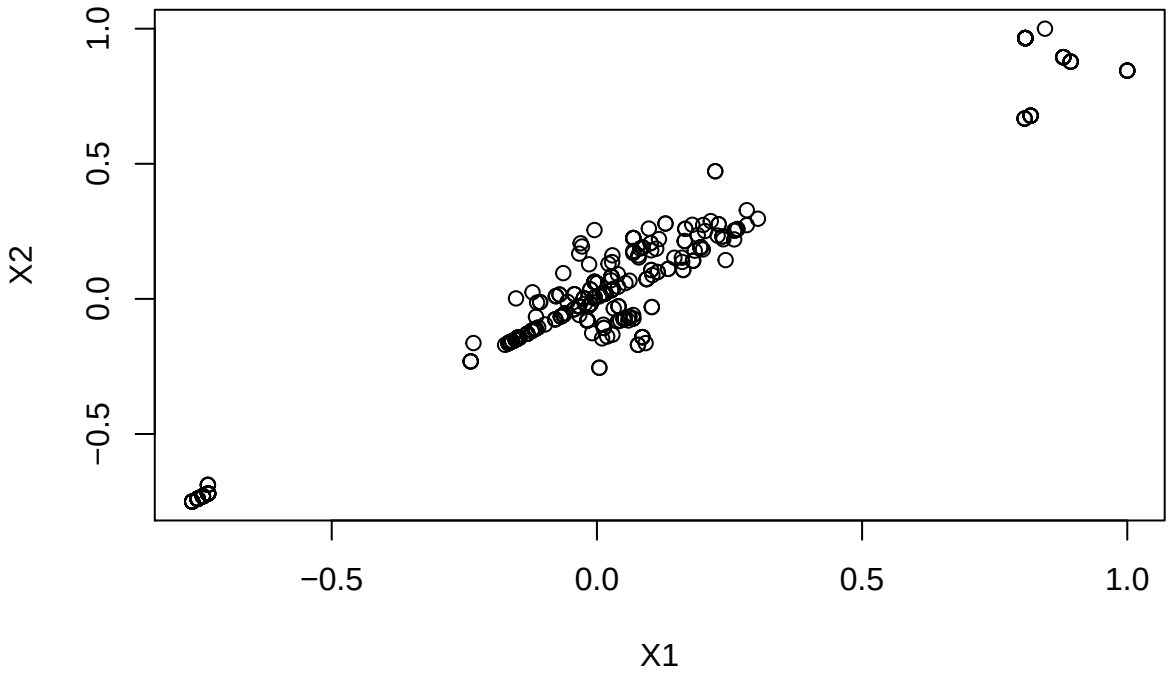
The fingerprint predictors indicate the presence or absence of substructures of a molecule and are often sparse meaning that relatively few of the molecules contain each substructure. Filter out the predictors that have low frequencies using the `nearZeroVar` function from the caret package. How many predictors are left for modeling?

```
remove.features <- nearZeroVar(fingerprints)
X <- fingerprints[,-remove.features]
length(remove.features) %>% paste('columns are removed. ', dim(X)[2], ' predictors are left for modelin
```

```
## [1] "719 columns are removed.  388  predictors are left for modeling."
```

We will now look into pairwise correlation above 0.90. We will then remove the predictors having correlation with cutoff 0.90.

```
plot(cor(X))
```




```
corr.90 <- findCorrelation(cor(X), cutoff=0.90)
X <- X[,-corr.90]
length(corr.90) %>% paste('columns having correlation 0.90 or more are removed. ', dim(X)[2], ' predicto
```

```
## [1] "278 columns having correlation 0.90 or more are removed.  110  predictors are left for modeling
```

(c)

Split the data into a training and a test set, pre-process the data, and tune a PLS model. How many latent variables are optimal and what is the corresponding resampled estimate of R^2?

```
set.seed(786)
partition <- createDataPartition(permeability, p=0.70, list = FALSE)

# predictor
X.train <- X[partition,]
X.test <- X[-partition,]

# response
y.train <- permeability[partition,]
y.test <- permeability[-partition,]

# tune pls model
pls.fit <- train(x=X.train,
                 y=y.train,
                 method="pls",
                 metric="Rsquared",
                 tuneLength=15,
                 trControl=trainControl(method = "cv", verboseIter = TRUE),
                 preProcess=c("center", "scale"),
                 verbose=TRUE
                 )
```

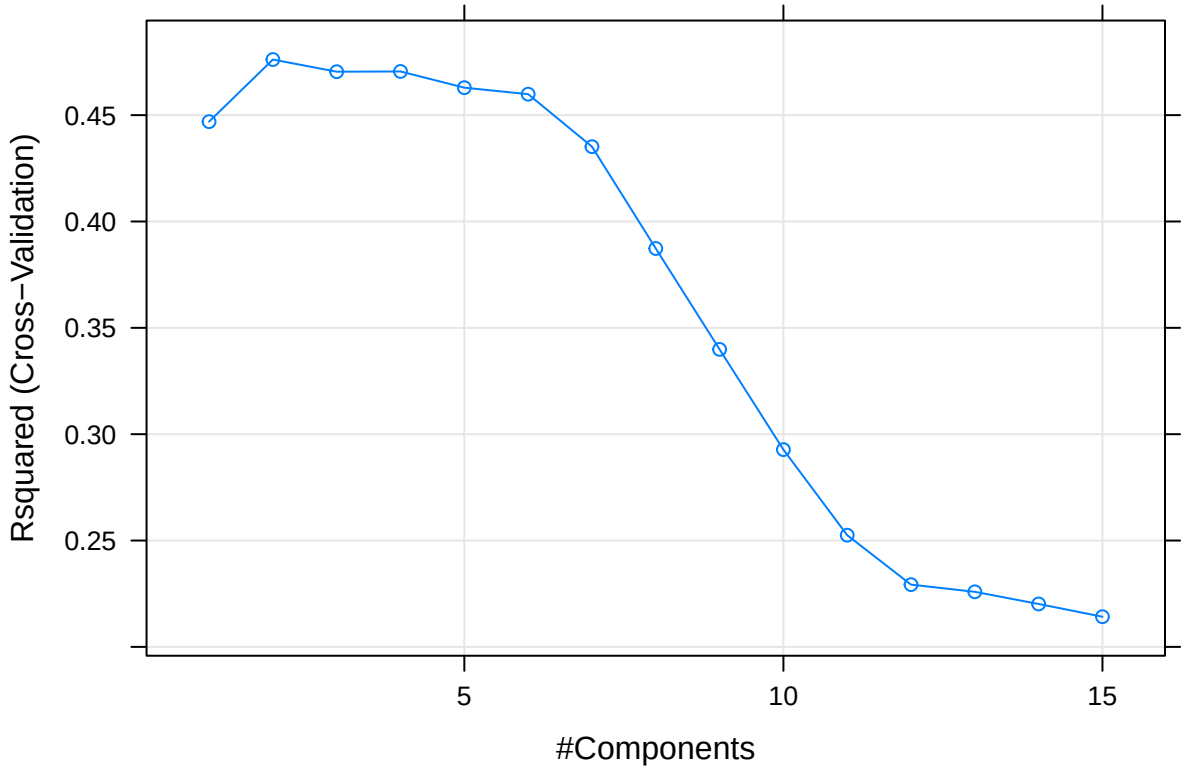
```
## + Fold01: ncomp=15
```

```
## - Fold01: ncomp=15
## + Fold02: ncomp=15
## - Fold02: ncomp=15
## + Fold03: ncomp=15
## - Fold03: ncomp=15
## + Fold04: ncomp=15
## - Fold04: ncomp=15
## + Fold05: ncomp=15
## - Fold05: ncomp=15
## + Fold06: ncomp=15
## - Fold06: ncomp=15
## + Fold07: ncomp=15
## - Fold07: ncomp=15
## + Fold08: ncomp=15
## - Fold08: ncomp=15
## + Fold09: ncomp=15
## - Fold09: ncomp=15
## + Fold10: ncomp=15
## - Fold10: ncomp=15
## Aggregating results
## Selecting tuning parameters
## Fitting ncomp = 2 on full training set
```

```
pls.fit
```

```
## Partial Least Squares
##
## 117 samples
## 110 predictors
##
## Pre-processing: centered (110), scaled (110)
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 107, 105, 105, 106, 105, 105, ...
## Resampling results across tuning parameters:
##
##   ncomp  RMSE      Rsquared   MAE
##    1     11.63701  0.4469410   8.212073
##    2     11.19413  0.4761420   8.071552
##    3     11.33762  0.4704264   8.606355
##    4     11.37860  0.4705414   8.654649
##    5     11.49895  0.4629129   8.619412
##    6     11.77805  0.4598600   8.916459
##    7     12.31715  0.4351858   9.354903
##    8     13.33637  0.3873041   9.915650
##    9     13.96776  0.3398617  10.369458
##   10     14.81892  0.2927483  10.927883
##   11     15.56533  0.2525077  11.412582
##   12     16.37035  0.2292636  11.999733
##   13     16.76865  0.2258902  12.449619
##   14     17.30096  0.2201694  12.898988
##   15     17.80399  0.2141806  13.206039
##
## Rsquared was used to select the optimal model using the largest value.
## The final value used for the model was ncomp = 2.
```

```
# plot R-squared vs components
plot(pls.fit)
```




```
pls.fit$results %>%
  filter(ncomp == pls.fit$bestTune$ncomp) %>%
  select(ncomp,RMSE,Rsquared)
```

```
##   ncomp     RMSE Rsquared
## 1     2 11.19413 0.476142
```

After applying partial least square model, we see now that number of components 2 produces minimum RMSE (11.19413) and max R^2 (0.476142).

(d)

Predict the response for the test set. What is the test set estimate of R^2?

(e)

Try building other models discussed in this chapter. Do any have better predictive performance?

(f)

Would you recommend any of your models to replace the permeability laboratory experiment?

Exercise 6.3

A chemical manufacturing process for a pharmaceutical product was discussed in Sect. 1.4. In this problem, the objective is to understand the relationship between biological measurements of the raw materials (predictors), measurements of the manufacturing process (predictors), and the response of product yield. Biological

predictors cannot be changed but can be used to assess the quality of the raw material before processing. On the other hand, the manufacturing process predictors can be changed in the manufacturing process. Improving product yield by 1% will boost revenue by approximately one hundred thousand dollars per batch:

(a)

Start R and use these commands to load the data:

The matrix `processPredictors` contains the 57 predictors (12 describing the input biological material and 45 describing the process predictors) for the 176 manufacturing runs. yield contains the percent yield for each run.

(b)

A small percentage of cells in the predictor set contain missing values. Use an imputation function to fill in these missing values (e.g., see Sect. 3.8).

(c)

Split the data into a training and a test set, pre-process the data, and tune a model of your choice from this chapter. What is the optimal value of the performance metric?

(d)

Predict the response for the test set.What is the value of the performance metric and how does this compare with the resampled performance metric on the training set?

(e)

Which predictors are most important in the model you have trained? Do either the biological or process predictors dominate the list?

(f)

Explore the relationships between each of the top predictors and the response. How could this information be helpful in improving yield in future runs of the manufacturing process?