

Data624 - Homework7

Amit Kapoor

3/28/2021

Contents

Exercise 6.2	1
(a)	2
(b)	2
(c)	2
(d)	2
(e)	2
(f)	3
Exercise 6.3	3
(a)	3
(b)	3
(c)	3
(d)	3
(e)	3
(f)	3

```
library(AppliedPredictiveModeling)
library(psych)
library(tidyverse)
```

```
## -- Attaching packages ----- tidyverse 1.3.0 --

## v ggplot2 3.3.3      v purrr  0.3.4
## v tibble  3.1.0      v dplyr  1.0.5
## v tidyr   1.1.3      v stringr 1.4.0
## v readr   1.4.0      v forcats 0.5.1

## -- Conflicts ----- tidyverse_conflicts() --
## x ggplot2::%+%( ) masks psych::%+%( )
## x ggplot2::alpha() masks psych::alpha()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag() masks stats::lag()
```

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?

(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 ?

(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?