ST563 - Homework 3

Wenbin Lu

1 Instructions

Please follow the instructions below when you prepare and submit your assignment.

- Include a cover page with your homework. It should contain
  1. Full name
  2. Course #
  3. HW #
  4. Submission date
- Assignments should be submitted using **Gradescope**.
- Neatly typed work should be submitted. All R code/output should be well commented on, with relevant outputs highlighted and discussed.
- When you solve a particular problem, do not only give the final answer. Instead, show your work and the steps you used (with proper explanation) to arrive at your answer to get full credit.
- Submission in the PDF format is preferred; Please convert other formats such as doc or docx into PDF.

Problem 1 (25 points)

Do Chapter 6, Problem 8

Problem 2 (25 points)

Do Chapter 6, Problem 9

Prolem 3 (25 points)

Problem adapted from Applied Predictive Modeling by Kuhn, and Johnson

> Infrared (IR) spectroscopy technology can be used to determine the chemical makeup of a substance. The theory of IR spectroscopy holds that unique molecular structures absorb IR frequencies differently. In practice a spectrometer fires a series of IR frequencies into a sample material, and the device measures the absorbance of the sample at each individual frequency. This series of measurements creates a spectrum profile which can then be used to determine the chemical make–up of the sample material.
>
> A Tecator Infratec Food and Feed Analyzer instrument was used to analyze 215 samples of meat across 100 frequencies. A sample of these frequency profiles are displayed in Figure 1. In addition an IR profile, analytical chemistry was used to determine the percent content of water, fat, and protein for each sample. If we can establish a predictive relationship between IR spectrum and fat content, then food scientists could predict a sample's fat content instead of using analytical chemistry to determine the content. This would provide costs savings, since analytical chemistry is a more expensive, time consuming process.

Load the data using following commands. See `?tecator` for details.

```
library(caret)
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'caret'
```

```
## The following object is masked from 'package:purrr':
##
##     lift
```

```
data(tecator)
```

You should now have access to two matrices:

- *absorp*: absorbance data for 215 samples. Each row is one observation, and each column is one predictor.
- *endpoints*: outcome variables. Measured are the percentages of water, fat and protein in the samples (columns 1, 2 and 3, resectively).

We want to predict *fat* content based on the absorbance data.

(a) In the absorbance data, the predictors are the measurements at the individual frequencies. Because the frequencies lie in a systematic order (850 – 1,050 nm, each column corresponds to one such frequency), the predictors have a high degree of correlation. Hence, the data lie in a smaller dimension than the total number of predictors (100). Plot the predictors as functions of frequencies (i.e, plot each row of the data as a line) in one plot. Explain if you see any patterns.

(b) Do you think the data is indeed 100-dimensional, or could we reduce the dimension to a lower number? Use PCA on the absorbance data and explore the results, addressing each of the points below.

- How many PCs would you like to keep?
- Examine the loadings of the PCs you would like to keep. Is there a nice interpretation of the loadings in terms of how they contribute to the corresponding PCs, or what patterns they are capturing?
- Do the PCs you want to keep correlate with the response?

Do **not** print pages after pages of output. Instead, present summary of your exploration using appropriate tables/figures.

(c) Investigate the following models using your choice of data splitting method (holdout, CV etc) and compare their predictive performance.

- Principal components regression with number of components chosen by 10-fold CV.
- Partial least squares regression with number of components chosen by 10-fold CV.

Note: as before, parameter tuning happens in `inner loop`, and you choice of data splitting method is outer loop.

(d) Investigate the loadings of the first PLS component. Can you interpret the loadings, and the corresponding PLS component is a nice way?

(e) Create a plot showing absolute value of correlation between the response of the PLS components. Explain any patterns you observe.

Problem 4 (25 points)

Do Chapter 4, Problem 4.