Week3, Assignment 1

Student Name

```r
library(caret)
library(AppliedPredictiveModeling)
# ... add additional libraries here ...

seed <- 123
```

Problem 3.1 (30 points)

Infrared (IR) spectroscopy technology is 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 makeup of the sample material.

A Tecator Infratec Food and Feed Analyzer instrument was used to analyze 215 samples of meat across 100 frequencies. In addition to an IR profile, analytical chemistry determined 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 with IR instead of using analytical chemistry. This would provide costs savings since analytical chemistry is a more expensive, time-consuming process

3.1.a Load the data

```r
# library(caret)
data(tecator)
# Use ?tecator for details
```

The matrix absorp contains the 100 absorbance values for the 215 samples, while matrix endpoints contain the percent of moisture, **fat**, and protein in columns 1–3, respectively.

3.1.b (5 points)

In this example, the predictors are the measurements at the individual frequencies. Because the frequencies lie in a systematic order (850–1,050 nm), the predictors have a high degree of correlation. Hence, the data lie in a smaller dimension than the total number of predictors (100). **Use PCA to determine the effective dimension of these data. What is the effective dimension(it's the number of principal components needed to explain the majority of the variance in the data)?**

```
# ... code ...
```

... answer ...

3.1.c (20 points)

Split the data into a training and a validation set using a resampling technique, pre-process the data by centering and scaling, and build linear regression, partial least squares, ridge regression, lasso regression and elastic net models described in this chapter. For those models with tuning parameters, what are the optimal values of the tuning parameter(s)?

```
# ... code ...
```

...Optional recap (required if optimized parameters are not output by code)..

3.1.d (3 points)

Which model has the best predictive ability using RMSE, MAE and R2 (on the training results) as metrics? Is any model significantly better or worse than the others?

```
# train_metrics <- resamples( ... )
# summary(train_metrics)
# dotplot(train_metrics)
```

... Answer ...

3.1.e (2 points)

Explain which model you would use for predicting the fat content of a sample.

... Answer ...

Problem 3.2 (30 points)

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:

3.2.a Load the data:

```
# library(AppliedPredictiveModeling)
data(permeability)
# use ?permeability to see more details
```

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

3.2.b (5 points)

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 after filtering?**

```
# ... code ...
```

... answer ...

3.2.c (5 points)

Split the data into a training (80%) and a test set (20%), pre-process the data, and tune a PLS model. How many latent variables are optimal, and what is the corresponding resampled estimate of R2?

```
# ... code ...
```

... Answer ...

3.2.d (3 points)

Predict the response for the test set. What is the test set estimate of R2?

```
# ... code ...
```

3.2.e (15 points)

Try building lasso, ridge and elastic net regression models discussed in this chapter. Do any have better predictive performance using R2 as metric on the test data?

```
# ... code ...
```

...Answer...

3.2.f (2 points)

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

...Answer...

Problem 3.3 (30 points)

A chemical manufacturing process for a pharmaceutical product was discussed in Section.1.4 of the textbook. 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, 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:

3.3.a

```
library(AppliedPredictiveModeling)
data(ChemicalManufacturingProcess)
```

The ChemicalManufacturingProcess data frame contains 57 predictors (12 describing the input biological material and 45 describing the process predictors) and a yield column which is the percent yield for each run for the 176 manufacturing runs.

3.3.b (4 points)

Split the data into a training (80%) and a test set (20%). A small percentage of cells in the predictor set contain missing values. Use an imputation function to fill in these missing values in both training and test data sets, also perform centering and scaling.

```
# ... code ...
```

…how many values were imputed?…

3.3.c (16 points)

Tune lasso regression model (lasso), ridge regression model (ridge), partial least squares model (pls), and elastic net model (enet) from chapter 6. What is the optimal value of the resampled performance metric RMSE?

```
# ... code ...
```

…Answer…

3.3.d (5 points)

Predict the response for the test set using the above trained models. What is the value of the performance metric RMSE, and how does this compare with the resampled performance metric RMSE on the training set?

```
# ... code ...
```

3.3.e (5 points)

In the optimal model, how many biological and process predictors remain (whose coefficient is greater than zero)? What are the top five predictors (use absolute values of coefficients) that have the most impact on the yield?

```
# ... code ...
```