HW7

Tianqi Wu

3/24/2020

```
library(readxl)
library(tidyverse)
library(glmnet)
library(caret)
library(compositions)
```

Problem 1

In one sentence, explain what the point of this study is

The study finds the relationship between composition of intestinal microbial communities and enteric infections and ways for prevention strategies.

In one sentence, report the major finding of this study.

The intestinal microbial communities of patients have more Proteobacteria representing genus Escherichia relative to communities of healhty family members, which were dominated by Bacteroides and Firmicutes.

What type of study is this: a) Cohort, b) Cross-sectional, c) Case-control

It is a case-control study.

What population were the data in this study sampled from?

The population were sampled from patients with enteric infections at four participating hospitals and the Michigan Department of Health and Human Services (MDHHS).

Problem 2

How many subjects were in your final merged dataset?

The dimension of final merged dataset is (267, 256). There are 267 subjects in the merged dataset.

How many OTUs did you remove, and which ones?

I removed 1 OTU (k___Bacteria;p___Proteobacteria;c___Alphaproteobacteria;o___Rhizobiales;f___Phyllobacteriaceae;g___Phyl

How many OTUs were in your final merged dataset?

There are 252 OUTs in the merged dataset.

```
Add_4 = read_excel('40168_2015_109_MOESM4_ESM.xlsx')
Add_5 = read_excel('40168_2015_109_MOESM5_ESM.xlsx')

colnames(Add_5)[1] = 'SampleID'
removed = c()
for (i in 2:ncol(Add_5)){
  if (var(Add_5[, i]) == 0){
    removed = append(removed, colnames(Add_5)[i])
  }
}
Add_5_new = Add_5[ , !(names(Add_5) %in% removed)]
merged = inner_join(Add_4, Add_5_new, by='SampleID')
```

Problem 3

estimated testing error

The estimated testing error is 0.8224294.

true testing error

The true testing error is 0.1791045.

How many OTUs are used in the final prediction rule?

There are 37 OTUs used in the final prediction rule.

Which OTU seems to have the biggest effect?

"k___Bacteria;p___Firmicutes;c___Clostridia;o___Clostridiales;f___Eubacteriaceae;g___Anaerofustis" has the biggest negative effect of 1773.992 on the prediction.

```
set.seed(123)
LASSO = function(data){
  train_idx <- sample(1:nrow(data), nrow(data) * 0.75)
  train <- merged[train_idx, ]
  test <- merged[-train_idx, ]
  lasso <- cv.glmnet(x = data.matrix(train[,5:ncol(train)]),
                     y = as.factor(train$Status),
                     nfolds = 5,
                     family = "binomial")

  ## estimated testing error from cross-validation
  print(lasso)
  cat('estimated testing error:', min(lasso$cvm),'\n')

  ## predictions
  pred <- predict(lasso,
                  newx = data.matrix(test[,5:ncol(test)]),
                  s = "lambda.min",
                  type = "class")

  ## true testing error
  cat('true testing error:', mean(as.character(pred) != as.character(test$Status)),'\n')

  mycoef = coef(lasso, s = "lambda.min")
  max_coef = max(abs(mycoef))
  max_index = which(abs(mycoef)==max_coef)
  cat('max_coef:', max_coef,'\n')
  cat('OTU with the biggest effect:',mycoef@Dimnames[[1]][max_index],'\n')
}
LASSO(merged)
```

```
##
## Call:  cv.glmnet(x = data.matrix(train[, 5:ncol(train)]), y = as.factor(train$Status),      nfolds =
##
## Measure: Binomial Deviance
##
##      Lambda Measure      SE Nonzero
## min 0.02266  0.8224 0.11077      37
## 1se 0.10516  0.9312 0.03808      10
## estimated testing error: 0.8224294
## true testing error: 0.1791045
## max_coef: 1773.992
## OTU with the biggest effect: k__Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Eubacteriac
```

Problem 4

estimated testing accuracy

The estimated testing accuracy of the optimal tuning parameter using 5-fold cross-validation is 0.8849312 with parameters (mtry = 127, splitrule = extratrees and min.node.size = 1).

true testing error

The true testing error is 0.1044776.

Which algorithm performed better, the lasso, or random forest?

Random forest is better ince it has lower testing error (0.1044776 < 0.1791045).

```
set.seed(123)
RF = function(data){
  train_idx <- sample(1:nrow(data), nrow(data) * 0.75)
  train <- merged[train_idx, ]
  test <- merged[-train_idx, ]

  train_small <- train[, c(2, 5:ncol(train))]
  test_small <- test[, c(2, 5:ncol(test))]
  rf <- train(Status ~ .,
              data = train_small,
              method = "ranger",
              trControl = trainControl(method = "cv",
                                       number = 5))


  ## estimated testing accuracy from cross-validation
  print(rf)

  ## predictions
  pred <- predict(rf, test_small)

  cat('true testing error:',mean(as.character(pred) != as.character(test$Status)))
}
RF(merged)
```

```
## Random Forest
##
## 200 samples
## 252 predictors
##   2 classes: 'Healthy', 'Patient'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 159, 161, 159, 160, 161
## Resampling results across tuning parameters:
##
##   mtry  splitrule   Accuracy   Kappa
##     2   gini        0.8000219  0.3265035
##     2   extratrees  0.7351188  0.0000000
##   127   gini        0.8645528  0.6552090
##   127   extratrees  0.8849312  0.7069344
##   252   gini        0.8646748  0.6475092
##   252   extratrees  0.8800532  0.6898816
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 127, splitrule = extratrees
##  and min.node.size = 1.
## true testing error: 0.1044776
```

Problem 5

LASSO

estimated testing error

The estimated testing error is 0.8638855.

true testing error

The true testing error is 0.1641791.

How many OTUs are used in the final prediction rule?

There are 18 OTUs used in the final prediction rule.

Which OTU seems to have the biggest effect?

"k___Bacteria;p___Bacteroidetes;c___Bacteroidia;o___Bacteroidales;f___Rikenellaceae;g___AF12" has the biggest negative effect of 110.98 on the prediction.

Random Forest

estimated testing accuracy

The estimated testing accuracy of the optimal tuning parameter using 5-fold cross-validation is 0.8901032 with parameters (mtry = 127, splitrule = extratrees and min.node.size = 1).

true testing error

The true testing error is 0.07462687.

Which algorithm performed better, the lasso with transformed, or random forest with transformed?

Random forest with transformed is better ince it has lower testing error (0.07462687 < 0.1641791).

Which algorithm was best?

Random forest with transformed is best since it has the lowest testing error.

```r
clr_merged = clr(merged[5:ncol(merged)])
new_merged = merged
new_merged[5:ncol(merged)] = clr_merged
LASSO(new_merged)
```

```
##
## Call:  cv.glmnet(x = data.matrix(train[, 5:ncol(train)]), y = as.factor(train$Status),      nfolds =
##
## Measure: Binomial Deviance
##
##     Lambda Measure      SE Nonzero
## min 0.0525  0.8639 0.11156      18
## 1se 0.1007  0.9713 0.05883       8
## estimated testing error: 0.8638855
## true testing error: 0.1641791
## max_coef: 110.98
## OTU with the biggest effect: k__Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Rikenel
```

```
RF(new_merged)
```

```
## Random Forest
##
## 200 samples
## 252 predictors
##   2 classes: 'Healthy', 'Patient'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 160, 159, 160, 161, 160
## Resampling results across tuning parameters:
##
##   mtry  splitrule   Accuracy   Kappa
##     2   gini        0.7751939  0.2242499
##     2   extratrees  0.7300594  0.0000000
##   127   gini        0.8698468  0.6590798
##   127   extratrees  0.8901032  0.7086537
##   252   gini        0.8698405  0.6596356
##   252   extratrees  0.8849750  0.6903133
##
## Tuning parameter 'min.node.size' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were mtry = 127, splitrule = extratrees
##  and min.node.size = 1.
## true testing error: 0.07462687
```