Homework 4

Shuhei Kaneko

2022-11-02

```
library(tidyverse)
library(tidymodels)
library(MASS)
library(discrim)
```

Question 1

```
titanic <- read_csv("titanic.csv")
survived_reorder <- factor(titanic$survived,
                           levels = c("Yes", "No"))
titanic <- titanic %>%
  mutate(survived = survived_reorder) %>%
  mutate(pclass = as.factor(pclass))

# split & stratify the data
set.seed(10)

titanic_split <- initial_split(titanic, prop = 0.80,
                                strata = survived)
titanic_train <- training(titanic_split)
titanic_test <- testing(titanic_split)
```

Check the dimension of training and testing data.

```
dim(titanic_train)
```

```
## [1] 712  12
```

```
dim(titanic_test)
```

```
## [1] 179  12
```

Question 2

```
titanic_folds <- vfold_cv(titanic_train, v = 10)
titanic_folds
```

```
## #  10-fold cross-validation
## # A tibble: 10 x 2
##    splits           id
##    <list>           <chr>
##  1 <split [640/72]> Fold01
##  2 <split [640/72]> Fold02
##  3 <split [641/71]> Fold03
```

```
##  4 <split [641/71]> Fold04
##  5 <split [641/71]> Fold05
##  6 <split [641/71]> Fold06
##  7 <split [641/71]> Fold07
##  8 <split [641/71]> Fold08
##  9 <split [641/71]> Fold09
## 10 <split [641/71]> Fold10
```

Question 3

In the previous question, the entire training data were randomly divided into 10 groups with equal size.

K-Hold cross validation: (K-1)/K of the sample will be used for training and the remained 1/K of the sample will be used for assessment.

If we simply fit and test its performance on the entire set, the resulting model would be overfitting the training data. By employing K-hold cross validation, we can discount the performance of the overfitted model.

If we did use the entire training set, this method would be called hold-out method.

Question 4

Recipe is identical with Homework 3

```
titanic_recipe <- recipe(survived ~ pclass + sex + age + sib_sp + parch + fare, data = titanic_train) %>
  step_impute_linear(age) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("sex"):fare) %>%
  step_interact(terms = ~ age:fare)
```

Set up the workflow

(i) Logistic regression

```
log_reg <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")

# Workflow
log_wkflow <- workflow() %>%
  add_model(log_reg) %>%
  add_recipe(titanic_recipe)
```

(ii) Linear discriminant analysis

```
# Engine
lda <- discrim_linear() %>%
  set_engine("MASS") %>%
  set_mode("classification")

# Workflow
lda_wkflow <- workflow() %>%
  add_model(lda) %>%
  add_recipe(titanic_recipe)
```

(iii) Quadratic discriminant analysis

```
# Engine
qda <- discrim_quad() %>%
```

```r
  set_engine("MASS") %>%
  set_mode("classification")

# Workflow
qda_wkflow <- workflow() %>%
  add_model(qda) %>%
  add_recipe(titanic_recipe)
```

(Number of models) * (Number of Folds) = 3 * 10 = 30

We fit 30 models to the training data.

Question 5

```r
tune_log <- tune_grid(object = log_wkflow,
  resamples = titanic_folds)

tune_lda <- tune_grid(object = lda_wkflow,
  resamples = titanic_folds)

tune_qda <- tune_grid(object = qda_wkflow,
  resamples = titanic_folds)

save(tune_log, tune_lda, tune_qda,
     file = "k_fold_cv.rda")
```

```r
load(file = "k_fold_cv.rda")
```

Question 6

```r
collect_metrics(tune_log)
```

```
## # A tibble: 2 x 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.812    10  0.0137 Preprocessor1_Model1
## 2 roc_auc  binary     0.849    10  0.0153 Preprocessor1_Model1
```

```r
collect_metrics(tune_lda)
```

```
## # A tibble: 2 x 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.798    10  0.0136 Preprocessor1_Model1
## 2 roc_auc  binary     0.849    10  0.0146 Preprocessor1_Model1
```

```r
collect_metrics(tune_qda)
```

```
## # A tibble: 2 x 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.785    10  0.0103 Preprocessor1_Model1
## 2 roc_auc  binary     0.832    10  0.0155 Preprocessor1_Model1
```

Logistic regression has the highest mean prediction accuracy. The standard error of the accuracy of logistic regression and LDA is pretty similar, while that of QDA is slightly lower.

It implies that QDA is superior in terms of the variance of prediction accuracy. However, compared to the difference of mean accuracy (between logistic and QDA), the difference in standard error is very tiny. Judging from the above observations, I decide that logistic regression is the best model in this case.

Question 7

```
log_fit_alltrain <- fit(log_wkflow, titanic_train)

log_acc <- augment(log_fit_alltrain, new_data = titanic_train) %>%
  accuracy(truth = survived, estimate = .pred_class)

log_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.823
```

Question 8

```
pred_log_test <- predict(log_fit_alltrain, new_data = titanic_test, type = "prob")

log_test_acc <- augment(log_fit_alltrain, new_data = titanic_test) %>%
  accuracy(truth = survived, estimate = .pred_class)

log_test_acc
```

```
## # A tibble: 1 x 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.771
```

Question 9

OLS estimate of β can be derived by solving the following minimization problem:

$$\min \sum_{i=1}^{n}(Y_i - \beta)^2$$

By taking first order derivative wrt β, we can derive that

$$\frac{1}{n} * \sum_{i=1}^{n} Y_i - \hat{\beta} = 0\hat{\beta} = \frac{1}{n} * \sum_{i=1}^{n} Y_i = \bar{Y}$$

Question 10 Using the formula derived in the previous question, we can find

$$\hat{\beta}^{(1)} = \frac{1}{n-1}(Y_2 + Y_3 + \cdots, Y_n)\hat{\beta}^{(2)} = \frac{1}{n-1}(Y_1 + Y_3 + \cdots, Y_n)$$

The covariance of the above two is:

$$\begin{aligned} Cov(\hat{\beta}^{(1)}, \hat{\beta}^{(2)}) &= Cov(\frac{1}{n-1}(Y_2 + Y_3 + \cdots, Y_n), \frac{1}{n-1}(Y_1 + Y_3 + \cdots, Y_n)) \\ &= \frac{n-2}{(n-1)^2}\sigma^2 \end{aligned}$$