131_HW4

Immad Ali

5/3/2022

Resampling

For this assignment, we will continue working with part of a Kaggle data set (https://www.kaggle.com/c/titanic/overview) that was the subject of a machine learning competition and is often used for practicing ML models. The goal is classification; specifically, to predict which passengers would survive the Titanic shipwreck (https://en.wikipedia.org/wiki/Titanic).

Question 1

Split the data, stratifying on the outcome variable, `survived.` You should choose the proportions to split the data into. Verify that the training and testing data sets have the appropriate number of observations.

```
set.seed(999)

#Set proportion to 60%
titanic_split <- initial_split(titanic, prop = 0.60, strata = survived)
titanic_train <- training(titanic_split)

titanic_test <- testing(titanic_split)
```

```
#checks nomber of observations are correct
dim(titanic_train)
```

```
## [1] 534  12
```

```
dim(titanic_test)
```

```
## [1] 357  12
```

The data matches up with the number of observations before the split.

Question 2

Fold the **training** data. Use *k*-fold cross-validation, with $k = 10$.

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

```
## #  10-fold cross-validation
## # A tibble: 10 × 2
##    splits           id
##    <list>           <chr>
##  1 <split [480/54]> Fold01
##  2 <split [480/54]> Fold02
##  3 <split [480/54]> Fold03
##  4 <split [480/54]> Fold04
##  5 <split [481/53]> Fold05
##  6 <split [481/53]> Fold06
##  7 <split [481/53]> Fold07
##  8 <split [481/53]> Fold08
##  9 <split [481/53]> Fold09
## 10 <split [481/53]> Fold10
```

Question 3

In your own words, explain what we are doing in Question 2. What is *k*-fold cross-validation? Why should we use it, rather than simply fitting and testing models on the entire training set? If we **did** use the entire training set, what resampling method would that be?

In Q2, we divided up the training data set into 10 groups (folds). K fold cross validation is a technique used to determine a models performance by splitting the trainign data set into *k* folds. IF we used the entire training data set, then we would use validation set approach.

Question 4

Set up workflows for 3 models:

1. A logistic regression with the `glm` engine;
2. A linear discriminant analysis with the `MASS` engine;
3. A quadratic discriminant analysis with the `MASS` engine.

```
#First we make the recipe
titanic_recipe <- recipe(survived ~ pclass + sex + age + sib_sp + parch + fare, data = t
itanic_train) %>%
  step_impute_linear(age) %>%
  step_dummy(all_nominal_predictors()) %>%
  step_interact(terms = ~ starts_with("sex"):fare) %>%
  step_interact(~ age:fare)
```

```
# The first workflow is logistic regression using the GLM engine
log_reg <- logistic_reg() %>%
  set_engine("glm") %>%
  set_mode("classification")

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

```r
# The second workflow is Linear discriminant analysis (LDA)
lda_mod <- discrim_linear() %>%
  set_engine("MASS") %>%
  set_mode("classification")

lda_wkflow <- workflow() %>%
  add_model(lda_mod) %>%
  add_recipe(titanic_recipe)
```

```r
# The final workflow is quadratic discriminant analysis (QDA)
qda_mod <- discrim_quad() %>%
  set_mode("classification") %>%
  set_engine("MASS")

qda_wkflow <- workflow() %>%
  add_model(qda_mod) %>%
  add_recipe(titanic_recipe)
```

How many models, total, across all folds, will you be fitting to the data? To answer, think about how many folds there are, and how many models you'll fit to each fold.

Since we have 10 folds and 3 workflows, we will have 30 models for our data.

Question 5

Fit each of the models created in Question 4 to the folded data.

```r
#Fitting each model
log_fit <- log_wkflow %>%
  fit_resamples(titanic_folds)

lda_fit <- lda_wkflow %>%
  fit_resamples(titanic_folds)

qda_fit <- qda_wkflow %>%
  fit_resamples(titanic_folds)
```

Question 6

Use `collect_metrics()` to print the mean and standard errors of the performance metric *accuracy* across all folds for each of the four models.

Decide which of the 3 fitted models has performed the best. Explain why. *(Note: You should consider both the mean accuracy and its standard error.)*

```r
collect_metrics(log_fit)
```

```
## # A tibble: 2 × 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.794    10  0.0170 Preprocessor1_Model1
## 2 roc_auc  binary     0.850    10  0.0140 Preprocessor1_Model1
```

```
collect_metrics(lda_fit)
```

```
## # A tibble: 2 × 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.774    10  0.0198 Preprocessor1_Model1
## 2 roc_auc  binary     0.853    10  0.0143 Preprocessor1_Model1
```

```
collect_metrics(qda_fit)
```

```
## # A tibble: 2 × 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.766    10  0.0167 Preprocessor1_Model1
## 2 roc_auc  binary     0.846    10  0.0202 Preprocessor1_Model1
```

Based on the metrics shown above, Logistic Regression seems to be the best fit for our data as it hads the lowest standard error and highest mean accuracy.

Question 7

Now that you've chosen a model, fit your chosen model to the entire training dataset (not to the folds).

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

Question 8

Finally, with your fitted model, use `predict()` , `bind_cols()` , and `accuracy()` to assess your model's performance on the testing data!

Compare your model's testing accuracy to its average accuracy across folds. Describe what you see.

```
log_predict <- predict(lm_fit, new_data = titanic_train, type = "prob")
log_predict <- bind_cols(log_predict, titanic_train %>% select(survived))

augment(lm_fit, new_data = titanic_train) %>%
  accuracy(as.factor(survived), estimate = .pred_class)
```

```
## # A tibble: 1 × 3
##   .metric  .estimator .estimate
##   <chr>    <chr>          <dbl>
## 1 accuracy binary         0.807
```

According to the field, my models testing accuracy is .807. The average accuracy across my folds was .794. This means our model’s testing accuracy is better than the average accuracy from 10 folds.