week11

Chang Dong

2023-04-13

```
pacman::p_load(tidymodels, tidyverse,titanic,ggplot2,discrim)
```

Linear Discriminant Analysis (LDA)

You are going to fit LDA on the Titanic dataset to see how it compares to the logistic regression you fit in Week 8: Example — Titanic and Yardstick. We used the following code:

```
library( titanic )
titanic <- as_tibble( titanic::titanic_train )
titanic <- titanic %>% mutate(Survived = factor( Survived ),
                              Pclass = factor( Pclass ),
                              Sex = factor( Sex )) %>%
  dplyr::select(Survived, Pclass, Age, Sex) %>%
drop_na()
```

What you need to do:

1. Fit LDA on the Titanic dataset using TidyModels. Remember, you are looking to classify whether people survived based on their age, sex, and class.

```
titanic_lda <- discrim_linear( mode = "classification" ) %>%
  set_engine( "MASS" ) %>%
  fit( Survived ~ Age + Sex + Pclass, data = titanic )
titanic_lda
```

```
## parsnip model object
##
## Call:
## lda(Survived ~ Age + Sex + Pclass, data = data)
##
## Prior probabilities of groups:
##         0         1
## 0.5938375 0.4061625
##
## Group means:
##        Age   Sexmale   Pclass2   Pclass3
## 0 30.62618 0.8490566 0.2122642 0.6367925
## 1 28.34369 0.3206897 0.2862069 0.2931034
##
## Coefficients of linear discriminants:
##                 LD1
## Age     -0.02275996
## Sexmale -1.99858372
## Pclass2 -0.86598339
## Pclass3 -1.69496397
```

2. Obtain class predictions for the titanic dataset.

```
titanic_preds <- predict( titanic_lda, new_data = titanic ) %>%
  bind_cols( titanic %>%
dplyr::select( Survived ) )
```

3. Get the confusion matrix for this model.

```
titanic_preds %>%
  conf_mat( Survived, .pred_class )
```

```
##           Truth
## Prediction   0   1
##          0 361  86
##          1  63 204
```

4. What is the sensitivity of this model? What is the specificity?

```
tibble( Sensitivity = 361 / (361 + 63),
        Specificity = 204 / (204 + 86) )
```

```
## # A tibble: 1 × 2
##   Sensitivity Specificity
##         <dbl>       <dbl>
## 1       0.851       0.703
```

5. Obtain probability predictions for this model.

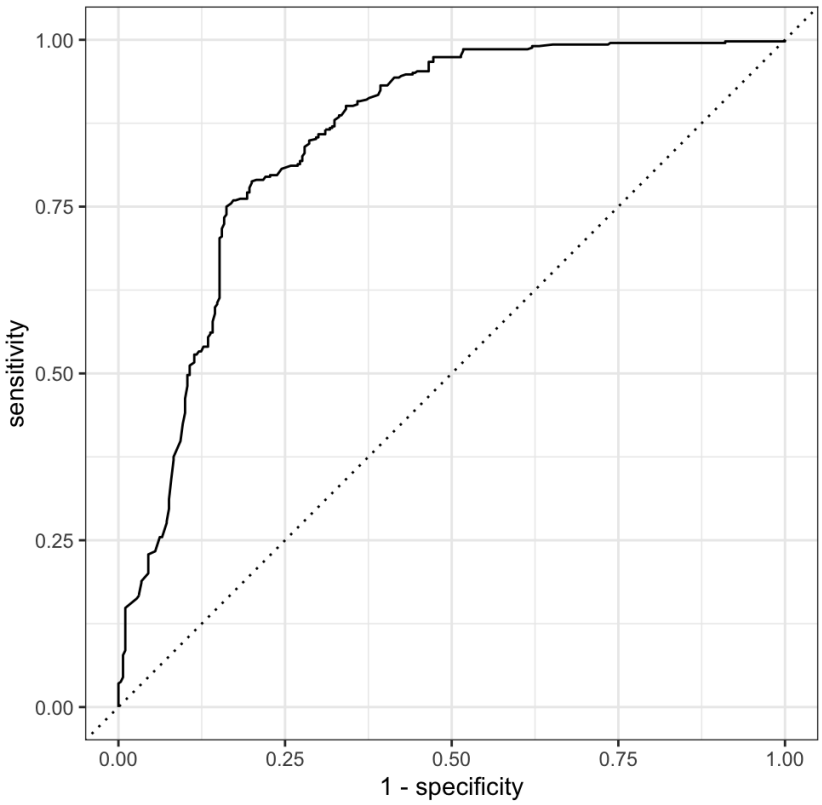
```
titanic_preds <- titanic_preds %>%
  bind_cols(predict( titanic_lda,
                     new_data = titanic,
                     type = "prob" )
)
titanic_preds
```

```
## # A tibble: 714 × 4
##    .pred_class Survived .pred_0 .pred_1
##    <fct>       <fct>      <dbl>   <dbl>
##  1 0           0          0.929   0.0707
##  2 1           1          0.0553  0.945
##  3 1           1          0.371   0.629
##  4 1           1          0.0497  0.950
##  5 0           0          0.955   0.0449
##  6 0           0          0.732   0.268
##  7 0           0          0.862   0.138
##  8 1           1          0.380   0.620
##  9 1           1          0.0896  0.910
## 10 1           1          0.207   0.793
## # … with 704 more rows
```

6. Plot the ROC curve for this model.

```
titanic_preds %>% roc_curve( truth = Survived,
                             estimate = .pred_0 ) %>%
  autoplot()
```

```
## Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
## dplyr 1.1.0.
## ℹ Please use `reframe()` instead.
## ℹ When switching from `summarise()` to `reframe()`, remember that `reframe()`
##   always returns an ungrouped data frame and adjust accordingly.
## ℹ The deprecated feature was likely used in the yardstick package.
##   Please report the issue at < ]8;;https://github.com/tidymodels/yardstick/issues https://github.co
m/tidymodels/yardstick/issues ]8;; >.
```



7. Obtain the AUC for this ROC curve.

```
titanic_preds %>%
  roc_auc(truth = Survived, estimate = .pred_0)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.850
```

8. Based on this value, would you prefer the logistic regression model or the LDA model to model this data?

The AUC for logistic regression is 0.852 > 0.850, so you would only just prefer the logistic regression. It is also much easier to interpret logistic regression, so this makes it even more preferred.

9. Create a 10-fold cross-validation set for this data (strata by Survived).

```
titanic_cv <- vfold_cv( titanic, v = 10, strata = Survived )
titanic_cv
```

```
## #  10-fold cross-validation using stratification
## # A tibble: 10 × 2
##    splits           id
##    <list>           <chr>
##  1 <split [642/72]> Fold01
##  2 <split [642/72]> Fold02
##  3 <split [642/72]> Fold03
##  4 <split [642/72]> Fold04
##  5 <split [643/71]> Fold05
##  6 <split [643/71]> Fold06
##  7 <split [643/71]> Fold07
##  8 <split [643/71]> Fold08
##  9 <split [643/71]> Fold09
## 10 <split [643/71]> Fold10
```

10. Using the cross-validation sets, do the following:

a. Create a model specification for LDA.

b. Fit LDA on the cross-validation sets.

c. Obtain the estimates of the AUC and accuracy, as well as their standard errors.

```
# a)
lda_spec <- discrim_linear( mode = "classification" ) %>%
  set_engine( "MASS" )

# b)
titanic_lda_resamples <- fit_resamples(
  object = lda_spec, preprocessor = recipe(Survived ~ . , data = titanic),
  resamples = titanic_cv)

# c)
titanic_lda_resamples %>%
  collect_metrics()
```

```
## # A tibble: 2 × 6
##   .metric  .estimator  mean     n std_err .config
##   <chr>    <chr>      <dbl> <int>   <dbl> <chr>
## 1 accuracy binary     0.780    10  0.0198 Preprocessor1_Model1
## 2 roc_auc  binary     0.852    10  0.0207 Preprocessor1_Model1
```