BIOST 546 Final Project

My-Anh Doan

2023-01-25

```
# set global options for code chunks
knitr::opts_chunk$set(message = FALSE, warning = FALSE, collapse = TRUE)
knitr::opts_knit$set(root.dir = rprojroot::find_rstudio_root_file())

library(dplyr)
library(knitr)
library(ggplot2)
library(caret)
```

```
# load data
load("./dataset/ADProj.RData")
str(ADProj, max.level = 1)
## List of 3
##  $ X_train: tibble [400 x 360] (S3: tbl_df/tbl/data.frame)
##  $ y_train: tibble [400 x 1] (S3: tbl_df/tbl/data.frame)
##  $ X_test : tibble [400 x 360] (S3: tbl_df/tbl/data.frame)

X_train <- ADProj[[1]]
y_train <- ADProj[[2]]
X_test <- ADProj[[3]]

train_dat <- data.frame(X_train, y_train)
contrasts(train_dat$Outcome)
##    AD
## C   0
## AD  1

# check for missing values
which(complete.cases(train_dat) == FALSE)
## integer(0)

# how many observations in each class in the training data
kable(train_dat %>% count(Outcome), caption = "# of observations in training diagnosis outcomes")
```

Table 1: # of observations in training diagnosis outcomes

Outcome	n
C	97
AD	303

The training data set above contains $n = 400$ observations and $p = 360$ predictors/features. In the training data set, there number of observations for each diagnosis class ("C" or "AD") are listed in Table 1.

```
glm_model <- glm(formula = Outcome ~ .,
                 family = binomial(link = "logit"),
                 data = train_dat)

#as.data.frame(summary(glm_model)$coefficients)

# training data
glm_prob_train <- predict(glm_model, type = "response", train_dat)
glm_label_train <- ifelse(glm_prob_train > 0.5, "AD", "C")

glm_train_matrix <- confusionMatrix(factor(glm_label_train, levels = c("AD","C")),
                                    factor(train_dat$Outcome, levels = c("AD","C")),
                                    positive = "AD")
kable(glm_train_matrix$table)
```

	AD	C
AD	303	0
C	0	97

```
glm_acc <- glm_train_matrix$overall[1]
glm_acc
## Accuracy
##        1

# test data
glm_prob_test <- predict(glm_model, type = "response", X_test)
glm_label_test <- ifelse(glm_prob_test > 0.5, "AD", "C")

test_counts <- glm_label_test %>%
  factor() %>%
  as.data.frame() %>%
  rename(Outcome = 1) %>%
  count(Outcome)

kable(test_counts, caption = "glm model: # of observations in test diagnosis outcomes")
```

Table 3: glm model: # of observations in test diagnosis outcomes

Outcome	n
AD	242
C	158

```
write.table(glm_label_test, "./FinalProj/DoanM_Pred1.txt",
            row.names = FALSE, col.names = FALSE, quote = FALSE)
```