Data Exploration

Andy Shen

10/27/2020

```
library(tidyverse)
library(MASS) #lda, qda
library(class) #knn
```

```
setwd("/Users/andyshen/Desktop/Git/Stats-101C-F20/Midterm Project")
train <- read.csv("training.csv")
test <- read.csv("test.csv")
```

```
set.seed(5732)
samp <- sample(1:nrow(train), floor(0.8 * nrow(train)), replace = FALSE)
train1 <- train[samp, ]
test_train <- train[-samp, ]
```

FamilyMemberCount, RVIS_percentile, N_Missense, intolerant_pNull, Gene_age, pLOF_Zscore

VEST_score

LDA

```
lda.mod <- lda(
  class ~ FamilyMemberCount + RVIS_percentile + N_Missense +
    intolerant_pNull + Gene_age + pLOF_Zscore, data = train1
)
preds <- predict(lda.mod, test_train, type = "response")$posterior
preds <- apply(preds, 1, which.max) - 1
tbl <- table(preds, test_train$class)
ter <- sum(diag(tbl)) / sum(tbl)
tbl
```

```
##
## preds   0   1   2
##     0 559  21  22
##     1   2   5   0
##     2  10   4  13
```

Test error rate is 0.093.

QDA

```
qda.mod <- qda(
  class ~ FamilyMemberCount + RVIS_percentile + N_Missense +
    intolerant_pNull + Gene_age + pLOF_Zscore, data = train1
)
```

```r
preds <- predict(qda.mod, test_train, type = "response")$posterior
preds <- apply(preds, 1, which.max) - 1
tbl <- table(preds, test_train$class)
ter <- sum(diag(tbl)) / sum(tbl)
tbl
```

```
##
## preds   0   1   2
##     0 503  12  10
##     1  35  14  10
##     2  33   4  15
```

Test error rate is 0.164.