Biostat 626 Midterm 1 Jingyu Wang

2023-04-08

Load the packages

```
library(tidyverse)
library(e1071)
library(FNN)
library(MASS)
library(rpart)
library(randomForest)
library(ggplot2)
```

Read the data

```
train_dat <- read.table("training_data.txt", sep = "", header = TRUE)
test_dat <- read.table("test_data.txt", sep = "", header = TRUE)
```

EDA

```
# check training data shape
train_shape <- dim(train_dat)
cat("Dimensions of the train data:", train_shape[1], "rows and", train_shape[2], "columns\n")
```

```
## Dimensions of the train data: 7767 rows and 563 columns
```

```
# check test data shape
test_shape <- dim(test_dat)
cat("Dimensions of the train data:", test_shape[1], "rows and", test_shape[2], "columns\n")
```

```
## Dimensions of the train data: 3162 rows and 562 columns
```

```
# check missing values in training data
colSums(is.na(train_dat))
```

```
##  subject activity       F1       F2       F3       F4       F5       F6 
##        0        0        0        0        0        0        0        0 
##       F7       F8       F9      F10      F11      F12      F13      F14 
##        0        0        0        0        0        0        0        0 
##      F15      F16      F17      F18      F19      F20      F21      F22 
##        0        0        0        0        0        0        0        0 
##      F23      F24      F25      F26      F27      F28      F29      F30 
##        0        0        0        0        0        0        0        0 
##      F31      F32      F33      F34      F35      F36      F37      F38 
##        0        0        0        0        0        0        0        0 
##      F39      F40      F41      F42      F43      F44      F45      F46 
##        0        0        0        0        0        0        0        0 
##      F47      F48      F49      F50      F51      F52      F53      F54 
##        0        0        0        0        0        0        0        0 
##      F55      F56      F57      F58      F59      F60      F61      F62 
##        0        0        0        0        0        0        0        0 
##      F63      F64      F65      F66      F67      F68      F69      F70 
##        0        0        0        0        0        0        0        0 
##      F71      F72      F73      F74      F75      F76      F77      F78 
##        0        0        0        0        0        0        0        0 
##      F79      F80      F81      F82      F83      F84      F85      F86 
##        0        0        0        0        0        0        0        0 
##      F87      F88      F89      F90      F91      F92      F93      F94 
##        0        0        0        0        0        0        0        0 
##      F95      F96      F97      F98      F99     F100     F101     F102 
##        0        0        0        0        0        0        0        0 
##     F103     F104     F105     F106     F107     F108     F109     F110 
##        0        0        0        0        0        0        0        0 
##     F111     F112     F113     F114     F115     F116     F117     F118 
##        0        0        0        0        0        0        0        0 
##     F119     F120     F121     F122     F123     F124     F125     F126 
##        0        0        0        0        0        0        0        0 
##     F127     F128     F129     F130     F131     F132     F133     F134 
##        0        0        0        0        0        0        0        0 
##     F135     F136     F137     F138     F139     F140     F141     F142 
##        0        0        0        0        0        0        0        0 
##     F143     F144     F145     F146     F147     F148     F149     F150 
##        0        0        0        0        0        0        0        0 
##     F151     F152     F153     F154     F155     F156     F157     F158 
##        0        0        0        0        0        0        0        0 
##     F159     F160     F161     F162     F163     F164     F165     F166 
##        0        0        0        0        0        0        0        0 
##     F167     F168     F169     F170     F171     F172     F173     F174 
##        0        0        0        0        0        0        0        0 
##     F175     F176     F177     F178     F179     F180     F181     F182 
##        0        0        0        0        0        0        0        0 
##     F183     F184     F185     F186     F187     F188     F189     F190 
##        0        0        0        0        0        0        0        0 
##     F191     F192     F193     F194     F195     F196     F197     F198 
##        0        0        0        0        0        0        0        0 
##     F199     F200     F201     F202     F203     F204     F205     F206 
##        0        0        0        0        0        0        0        0 
##     F207     F208     F209     F210     F211     F212     F213     F214 
##        0        0        0        0        0        0        0        0 
##     F215     F216     F217     F218     F219     F220     F221     F222 
##        0        0        0        0        0        0        0        0 
##     F223     F224     F225     F226     F227     F228     F229     F230 
##        0        0        0        0        0        0        0        0 
##     F231     F232     F233     F234     F235     F236     F237     F238 
##        0        0        0        0        0        0        0        0 
##     F239     F240     F241     F242     F243     F244     F245     F246 
##        0        0        0        0        0        0        0        0 
##     F247     F248     F249     F250     F251     F252     F253     F254 
##        0        0        0        0        0        0        0        0 
##     F255     F256     F257     F258     F259     F260     F261     F262 
##        0        0        0        0        0        0        0        0 
##     F263     F264     F265     F266     F267     F268     F269     F270 
##        0        0        0        0        0        0        0        0 
##     F271     F272     F273     F274     F275     F276     F277     F278 
##        0        0        0        0        0        0        0        0 
##     F279     F280     F281     F282     F283     F284     F285     F286 
##        0        0        0        0        0        0        0        0 
##     F287     F288     F289     F290     F291     F292     F293     F294 
##        0        0        0        0        0        0        0        0 
##     F295     F296     F297     F298     F299     F300     F301     F302 
##        0        0        0        0        0        0        0        0 
##     F303     F304     F305     F306     F307     F308     F309     F310 
##        0        0        0        0        0        0        0        0 
##     F311     F312     F313     F314     F315     F316     F317     F318 
##        0        0        0        0        0        0        0        0 
```

```
##     F319     F320     F321     F322     F323     F324     F325     F326
##        0        0        0        0        0        0        0        0
##     F327     F328     F329     F330     F331     F332     F333     F334
##        0        0        0        0        0        0        0        0
##     F335     F336     F337     F338     F339     F340     F341     F342
##        0        0        0        0        0        0        0        0
##     F343     F344     F345     F346     F347     F348     F349     F350
##        0        0        0        0        0        0        0        0
##     F351     F352     F353     F354     F355     F356     F357     F358
##        0        0        0        0        0        0        0        0
##     F359     F360     F361     F362     F363     F364     F365     F366
##        0        0        0        0        0        0        0        0
##     F367     F368     F369     F370     F371     F372     F373     F374
##        0        0        0        0        0        0        0        0
##     F375     F376     F377     F378     F379     F380     F381     F382
##        0        0        0        0        0        0        0        0
##     F383     F384     F385     F386     F387     F388     F389     F390
##        0        0        0        0        0        0        0        0
##     F391     F392     F393     F394     F395     F396     F397     F398
##        0        0        0        0        0        0        0        0
##     F399     F400     F401     F402     F403     F404     F405     F406
##        0        0        0        0        0        0        0        0
##     F407     F408     F409     F410     F411     F412     F413     F414
##        0        0        0        0        0        0        0        0
##     F415     F416     F417     F418     F419     F420     F421     F422
##        0        0        0        0        0        0        0        0
##     F423     F424     F425     F426     F427     F428     F429     F430
##        0        0        0        0        0        0        0        0
##     F431     F432     F433     F434     F435     F436     F437     F438
##        0        0        0        0        0        0        0        0
##     F439     F440     F441     F442     F443     F444     F445     F446
##        0        0        0        0        0        0        0        0
##     F447     F448     F449     F450     F451     F452     F453     F454
##        0        0        0        0        0        0        0        0
##     F455     F456     F457     F458     F459     F460     F461     F462
##        0        0        0        0        0        0        0        0
##     F463     F464     F465     F466     F467     F468     F469     F470
##        0        0        0        0        0        0        0        0
##     F471     F472     F473     F474     F475     F476     F477     F478
##        0        0        0        0        0        0        0        0
##     F479     F480     F481     F482     F483     F484     F485     F486
##        0        0        0        0        0        0        0        0
##     F487     F488     F489     F490     F491     F492     F493     F494
##        0        0        0        0        0        0        0        0
##     F495     F496     F497     F498     F499     F500     F501     F502
##        0        0        0        0        0        0        0        0
##     F503     F504     F505     F506     F507     F508     F509     F510
##        0        0        0        0        0        0        0        0
##     F511     F512     F513     F514     F515     F516     F517     F518
##        0        0        0        0        0        0        0        0
##     F519     F520     F521     F522     F523     F524     F525     F526
##        0        0        0        0        0        0        0        0
##     F527     F528     F529     F530     F531     F532     F533     F534
##        0        0        0        0        0        0        0        0
##     F535     F536     F537     F538     F539     F540     F541     F542
##        0        0        0        0        0        0        0        0
##     F543     F544     F545     F546     F547     F548     F549     F550
##        0        0        0        0        0        0        0        0
##     F551     F552     F553     F554     F555     F556     F557     F558
##        0        0        0        0        0        0        0        0
##     F559     F560     F561
##        0        0        0
```

```
# check missing values in training data
colSums(is.na(test_dat))
```

```
## subject      F1      F2      F3      F4      F5      F6      F7      F8      F9
##       0       0       0       0       0       0       0       0       0       0
##     F10     F11     F12     F13     F14     F15     F16     F17     F18     F19
##       0       0       0       0       0       0       0       0       0       0
##     F20     F21     F22     F23     F24     F25     F26     F27     F28     F29
##       0       0       0       0       0       0       0       0       0       0
##     F30     F31     F32     F33     F34     F35     F36     F37     F38     F39
##       0       0       0       0       0       0       0       0       0       0
##     F40     F41     F42     F43     F44     F45     F46     F47     F48     F49
##       0       0       0       0       0       0       0       0       0       0
##     F50     F51     F52     F53     F54     F55     F56     F57     F58     F59
##       0       0       0       0       0       0       0       0       0       0
##     F60     F61     F62     F63     F64     F65     F66     F67     F68     F69
##       0       0       0       0       0       0       0       0       0       0
##     F70     F71     F72     F73     F74     F75     F76     F77     F78     F79
##       0       0       0       0       0       0       0       0       0       0
##     F80     F81     F82     F83     F84     F85     F86     F87     F88     F89
##       0       0       0       0       0       0       0       0       0       0
##     F90     F91     F92     F93     F94     F95     F96     F97     F98     F99
##       0       0       0       0       0       0       0       0       0       0
##    F100    F101    F102    F103    F104    F105    F106    F107    F108    F109
##       0       0       0       0       0       0       0       0       0       0
##    F110    F111    F112    F113    F114    F115    F116    F117    F118    F119
##       0       0       0       0       0       0       0       0       0       0
##    F120    F121    F122    F123    F124    F125    F126    F127    F128    F129
##       0       0       0       0       0       0       0       0       0       0
##    F130    F131    F132    F133    F134    F135    F136    F137    F138    F139
##       0       0       0       0       0       0       0       0       0       0
##    F140    F141    F142    F143    F144    F145    F146    F147    F148    F149
##       0       0       0       0       0       0       0       0       0       0
##    F150    F151    F152    F153    F154    F155    F156    F157    F158    F159
##       0       0       0       0       0       0       0       0       0       0
##    F160    F161    F162    F163    F164    F165    F166    F167    F168    F169
##       0       0       0       0       0       0       0       0       0       0
##    F170    F171    F172    F173    F174    F175    F176    F177    F178    F179
##       0       0       0       0       0       0       0       0       0       0
##    F180    F181    F182    F183    F184    F185    F186    F187    F188    F189
##       0       0       0       0       0       0       0       0       0       0
##    F190    F191    F192    F193    F194    F195    F196    F197    F198    F199
##       0       0       0       0       0       0       0       0       0       0
##    F200    F201    F202    F203    F204    F205    F206    F207    F208    F209
##       0       0       0       0       0       0       0       0       0       0
##    F210    F211    F212    F213    F214    F215    F216    F217    F218    F219
##       0       0       0       0       0       0       0       0       0       0
##    F220    F221    F222    F223    F224    F225    F226    F227    F228    F229
##       0       0       0       0       0       0       0       0       0       0
##    F230    F231    F232    F233    F234    F235    F236    F237    F238    F239
##       0       0       0       0       0       0       0       0       0       0
##    F240    F241    F242    F243    F244    F245    F246    F247    F248    F249
##       0       0       0       0       0       0       0       0       0       0
##    F250    F251    F252    F253    F254    F255    F256    F257    F258    F259
##       0       0       0       0       0       0       0       0       0       0
##    F260    F261    F262    F263    F264    F265    F266    F267    F268    F269
##       0       0       0       0       0       0       0       0       0       0
##    F270    F271    F272    F273    F274    F275    F276    F277    F278    F279
##       0       0       0       0       0       0       0       0       0       0
##    F280    F281    F282    F283    F284    F285    F286    F287    F288    F289
##       0       0       0       0       0       0       0       0       0       0
##    F290    F291    F292    F293    F294    F295    F296    F297    F298    F299
##       0       0       0       0       0       0       0       0       0       0
##    F300    F301    F302    F303    F304    F305    F306    F307    F308    F309
##       0       0       0       0       0       0       0       0       0       0
##    F310    F311    F312    F313    F314    F315    F316    F317    F318    F319
##       0       0       0       0       0       0       0       0       0       0
##    F320    F321    F322    F323    F324    F325    F326    F327    F328    F329
##       0       0       0       0       0       0       0       0       0       0
##    F330    F331    F332    F333    F334    F335    F336    F337    F338    F339
##       0       0       0       0       0       0       0       0       0       0
##    F340    F341    F342    F343    F344    F345    F346    F347    F348    F349
##       0       0       0       0       0       0       0       0       0       0
##    F350    F351    F352    F353    F354    F355    F356    F357    F358    F359
##       0       0       0       0       0       0       0       0       0       0
##    F360    F361    F362    F363    F364    F365    F366    F367    F368    F369
##       0       0       0       0       0       0       0       0       0       0
##    F370    F371    F372    F373    F374    F375    F376    F377    F378    F379
##       0       0       0       0       0       0       0       0       0       0
##    F380    F381    F382    F383    F384    F385    F386    F387    F388    F389
##       0       0       0       0       0       0       0       0       0       0
##    F390    F391    F392    F393    F394    F395    F396    F397    F398    F399
##       0       0       0       0       0       0       0       0       0       0
```

```
##     F400    F401    F402    F403    F404    F405    F406    F407    F408    F409
##        0       0       0       0       0       0       0       0       0       0
##     F410    F411    F412    F413    F414    F415    F416    F417    F418    F419
##        0       0       0       0       0       0       0       0       0       0
##     F420    F421    F422    F423    F424    F425    F426    F427    F428    F429
##        0       0       0       0       0       0       0       0       0       0
##     F430    F431    F432    F433    F434    F435    F436    F437    F438    F439
##        0       0       0       0       0       0       0       0       0       0
##     F440    F441    F442    F443    F444    F445    F446    F447    F448    F449
##        0       0       0       0       0       0       0       0       0       0
##     F450    F451    F452    F453    F454    F455    F456    F457    F458    F459
##        0       0       0       0       0       0       0       0       0       0
##     F460    F461    F462    F463    F464    F465    F466    F467    F468    F469
##        0       0       0       0       0       0       0       0       0       0
##     F470    F471    F472    F473    F474    F475    F476    F477    F478    F479
##        0       0       0       0       0       0       0       0       0       0
##     F480    F481    F482    F483    F484    F485    F486    F487    F488    F489
##        0       0       0       0       0       0       0       0       0       0
##     F490    F491    F492    F493    F494    F495    F496    F497    F498    F499
##        0       0       0       0       0       0       0       0       0       0
##     F500    F501    F502    F503    F504    F505    F506    F507    F508    F509
##        0       0       0       0       0       0       0       0       0       0
##     F510    F511    F512    F513    F514    F515    F516    F517    F518    F519
##        0       0       0       0       0       0       0       0       0       0
##     F520    F521    F522    F523    F524    F525    F526    F527    F528    F529
##        0       0       0       0       0       0       0       0       0       0
##     F530    F531    F532    F533    F534    F535    F536    F537    F538    F539
##        0       0       0       0       0       0       0       0       0       0
##     F540    F541    F542    F543    F544    F545    F546    F547    F548    F549
##        0       0       0       0       0       0       0       0       0       0
##     F550    F551    F552    F553    F554    F555    F556    F557    F558    F559
##        0       0       0       0       0       0       0       0       0       0
##     F560    F561
##        0       0
```

Data Pre-Processing

```
# adding two variables for future prediction
train_dat <- train_dat %>%
  mutate(type1 = ifelse(activity %in% c(1:3), 1, 0),
         type2 = ifelse(activity %in% c(7:12), 7, activity))
```

Build models

task1

```
# split data
set.seed(123)
index <- sample(1:nrow(train_dat), nrow(train_dat)*0.7)
train1 <- train_dat[index, -c(1:2,565)]
test1 <- train_dat[-index, -c(1:2,565)]
```

```
# svm
# fit the model on the training set
svm_mod_1 <- svm(as.factor(type1)~., data = train1)
# make predictions on the test data
svm_pred_1 <- predict(svm_mod_1, test1)
# computing model accuracy rate
svm_acc_1 <- mean(svm_pred_1 == test1$type1)
svm_acc_1
```

```
## [1] 0.998713
```

```
# knn
# fit the model on the training set and make predictions on the test data
knn_pred_1 <- knn(train = train1[,-1], cl = train1$type1, test = test1[,-1])
# computing model accuracy rate
knn_acc_1 <- mean(knn_pred_1 == test1$type1)
knn_acc_1
```

```
## [1] 0.999571
```

```
# lda
# fit the model on the training set
lda_mod_1 <- lda(as.factor(type1)~., data = train1)
# make predictions on the test data
lda_pred_1 <- predict(lda_mod_1, test1)$class
# computing model accuracy rate
lda_acc_1 <- mean(lda_pred_1 == test1$type1)
lda_acc_1
```

```
## [1] 0.999571
```

```
# descision tree
# fit the model on the training set
tree_mod_1 <- rpart(as.factor(type1)~., data = train1)
# make predictions on the test data
tree_pred_1 <- predict(tree_mod_1, test1, type = "class")
# computing model accuracy rate
tree_acc_1 <- mean(tree_pred_1 == test1$type1)
tree_acc_1
```

```
## [1] 0.990991
```

```
# compare different methods
acc_df_1 <- data.frame(Method = c("SVM", "KNN", "LDA","DT"),
                       Accuray = c(svm_acc_1, knn_acc_1, lda_acc_1, tree_acc_1))
acc_df_1
```

Method <chr>	Accuray <dbl>
SVM	0.998713
KNN	0.999571
LDA	0.999571
DT	0.990991

4 rows

The best algorithms of task1 are KNN and LDA.

Task2

```
# split data
set.seed(123)
index <- sample(1:nrow(train_dat), nrow(train_dat)*0.7)
train2 <- train_dat[index, -c(1:2,564)]
test2 <- train_dat[-index, -c(1:2,564)]
```

```
# svm
# fit the model on the training set
svm_mod_2 <- svm(as.factor(type2)~., data = train2)
# make predictions on the test data
svm_pred_2 <- predict(svm_mod_2, test2)
# computing model accuracy rate
svm_acc_2 <- mean(svm_pred_2 == test2$type2)
svm_acc_2
```

```
## [1] 0.973402
```

```
# knn
# fit the model on the training set and make predictions on the test data
knn_pred_2 <- knn(train = train2[,-1], cl = train2$type2, test = test2[,-1])
# computing model accuracy rate
knn_acc_2 <- mean(knn_pred_2 == test2$type2)
knn_acc_2
```

```
## [1] 0.988417
```

```
# lda
# fit the model on the training set
lda_mod_2 <- lda(as.factor(type2)~., data = train2)
# make predictions on the test data
lda_pred_2 <- predict(lda_mod_2, test2)$class
# computing model accuracy rate
lda_acc_2 <- mean(lda_pred_2 == test2$type2)
lda_acc_2
```

```
## [1] 0.97855
```

```
# decision tree
# fit the model on the training set
dt_mod_2 <- rpart(as.factor(type2)~., data = train2)
# make predictions on the test data
dt_pred_2 <- predict(dt_mod_2, test2, type = "class")
# computing model accuracy rate
dt_acc_2 <- mean(dt_pred_2 == test2$type2)
dt_acc_2
```

```
## [1] 0.8687259
```

```
# random forest
# fit the model on the training set
rf_mod_2 <- randomForest(as.factor(type2)~., data = train2, ntree = 500, importance = TRUE)
# make predictions on the test data
rf_pred_2 <- predict(rf_mod_2, test2)
# computing model accuracy rate
rf_acc_2 <- mean(rf_pred_2 == test2$type2)
rf_acc_2
```

```
## [1] 0.976834
```

```
# compare different methods
acc_df_2 <- data.frame(Method = c("SVM", "KNN", "LDA","DT","RF"),
                     Accuray = c(svm_acc_2, knn_acc_2, lda_acc_2, dt_acc_2, rf_acc_2))
acc_df_2
```

Method <chr>	Accuray <dbl>
SVM	0.9734020
KNN	0.9884170
LDA	0.9785500
DT	0.8687259
RF	0.9768340

5 rows

The best algorithm of task2 is also KNN, but random forest also performed very well.

Final algorithm

Task1

```
# fit the model on the training set
knn_train_pred1 <- knn(train = train_dat[,-c(1,2,564,565)], cl = train_dat$type1, test = train_dat[,-c(1,2,564,56
5)])

# computing model accuracy rate
tree_train_acc1 <- mean(knn_train_pred1 == train_dat$type1)
tree_train_acc1
```

```
## [1] 1
```

```
# make predictions on the train data
knn_test_pred1 <- knn(train = train_dat[,-c(1,2,564,565)], cl = train_dat$type1, test = test_dat[,-1])

# save the results as txt file
write.table(knn_test_pred1, "binary_746766.txt", sep = "\t", quote = F,
            col.names = FALSE, row.names = FALSE)
```

Task2

```
# fit the model on the training set
knn_train_pred2 <- knn(train = train_dat[,-c(1,2,564,565)], cl = train_dat$type2, test = train_dat[,-c(1,2,564,56
5)])

# make predictions on the train data
knn_test_pred2 <- knn(train = train_dat[,-c(1,2,564,565)], cl = train_dat$type2, test = test_dat[,-1])

# save the results as txt file
write.table(knn_test_pred2, "multiclass_7467.txt", sep = "\t", quote = F,
            col.names = FALSE, row.names = FALSE)
```

Task 2 Final model: Random Forest

```
# fit the model on the train dataset
rf_train_mod <- randomForest(as.factor(type2)~., data = train2, ntree = 500, importance = TRUE)

# make predictions on the train data
rf_train_pred2 <- predict(rf_train_mod, train2)

# make predictions on the test data
rf_test_pred2 <- predict(rf_train_mod, test_dat[,-1])

# save the results as txt file
write.table(rf_test_pred2, "multiclass_746766.txt", sep = "\t", quote = F,
            col.names = FALSE, row.names = FALSE)
```