Winequality\_White

Charlotte Bahr

6/27/2020

library(kerasR) #DNN  
library(keras)  
library(tensorflow)#DNN  
library(data.table)  
library(dataPreparation)  
library(e1071) #Naive Bayes  
library(randomForest) #random Forest  
library(class) #k-Nearest Neighbor   
library(kernlab) #Support Vector Machine  
library(mlbench)#contains Glass data  
library(keras) #contains MNIST data  
library(caret)  
library(mltest)  
library(dplyr)  
use\_session\_with\_seed(9123)

## Set session seed to 9123 (disabled GPU, CPU parallelism)

## DATA PRE-PROCESSING  
wq <- read.csv("~/Desktop/Bachelorarbeit/R/R\_Files/winequality-white.csv", sep=";")  
  
wq$quality[wq$quality > 8] <- "8"  
wq$quality[wq$quality < 4] <- "4"  
  
wq <- as.data.frame(wq)  
  
set.seed(1234)  
train\_index\_SC <- sample(1:nrow(wq), 0.7\*nrow(wq))  
test\_index\_SC <- setdiff (1:nrow(wq), train\_index\_SC)  
  
Train <- wq[train\_index\_SC,]  
Test <- wq[test\_index\_SC,]  
  
True\_Label <- Test$quality  
  
#prep for DNN  
  
X\_train <- Train %>%   
 select(-quality) %>%   
 scale()  
  
Y\_train <- to\_categorical(Train$quality)  
  
X\_test <- Test %>%   
 select(-quality) %>%   
 scale()  
  
Y\_test <- to\_categorical(Test$quality)  
  
##MODELS   
  
#k-Nearest Neighbor  
  
pc <- proc.time()  
model\_KNN <- knn(Train, Test, as.factor(Train$quality), k=58)  
print(proc.time() - pc)

## user system elapsed   
## 0.198 0.008 0.213

#Naive Bayes  
  
pc <- proc.time()   
model\_NB <- naiveBayes(as.factor(Train$quality) ~. , Train)  
print(proc.time() - pc)

## user system elapsed   
## 0.012 0.006 0.017

#Random Forest  
  
pc <- proc.time()   
model\_RF <- randomForest(as.factor(Train$quality) ~. , Train)  
print(proc.time() - pc)

## user system elapsed   
## 2.997 0.125 3.125

#Support Vector Machine  
   
pc <- proc.time()   
model\_SVM <- ksvm(Train$quality ~. , Train, type = "C-svc", C = 1, kernel = "rbfdot" )  
print(proc.time() - pc)

## user system elapsed   
## 1.640 0.281 1.926

#Deep Neural Network  
  
  
pc <- proc.time()   
set.seed(42)  
  
model\_DNN <- Sequential()  
  
model\_DNN$add(Dense(units=250, input\_shape = dim(X\_train)[2]))  
model\_DNN$add(LeakyReLU())  
model\_DNN$add(Dropout(0.4))  
  
model\_DNN$add(Dense(units=250))  
model\_DNN$add(LeakyReLU())  
model\_DNN$add(Dropout(0.3))  
  
model\_DNN$add(Dense(units=250))  
model\_DNN$add(LeakyReLU())  
model\_DNN$add(Dropout(0.2))  
  
model\_DNN$add(Dense(9))  
model\_DNN$add(Activation("softmax"))  
  
# compile  
keras\_compile(model\_DNN, loss ="categorical\_crossentropy", optimizer = RMSprop(), metrics = "accuracy")  
keras\_fit(model\_DNN, X\_train, Y\_train, batch\_size = 128, epochs = 32, verbose= 1, validation\_split = 0.2)  
  
print(proc.time() - pc)

## user system elapsed   
## 6.056 0.457 6.492

##EVALUATION METRICS  
  
#predictions  
  
pred\_KNN <- #has no prediction value  
  
pred\_NB <- as.factor(predict(model\_NB, Test))  
  
pred\_RF <- as.factor(predict(model\_RF, Test))  
  
pred\_SVM <- as.factor(predict(model\_SVM, Test))  
  
pred\_DNN <- as.factor(keras\_predict\_classes(model\_DNN, X\_test))  
  
  
CF\_KNN <- table(model\_KNN, True\_Label)  
   
CF\_NB <- table(pred\_NB, True\_Label)  
   
CF\_RF <- table(pred\_RF, True\_Label)  
   
CF\_SVM <- table(pred\_SVM, True\_Label)  
  
CF\_DNN <- table(pred\_DNN, True\_Label)  
  
  
print(CF\_KNN)

## True\_Label  
## model\_KNN 4 5 6 7 8  
## 4 0 0 0 0 0  
## 5 38 161 141 27 6  
## 6 35 253 516 230 40  
## 7 0 0 12 6 5  
## 8 0 0 0 0 0

print(CF\_NB)

## True\_Label  
## pred\_NB 4 5 6 7 8  
## 4 14 14 10 1 0  
## 5 27 228 179 23 3  
## 6 17 127 226 58 9  
## 7 14 44 250 175 35  
## 8 1 1 4 6 4

print(CF\_RF)

## True\_Label  
## pred\_RF 4 5 6 7 8  
## 4 12 7 1 0 0  
## 5 34 286 93 8 0  
## 6 26 116 529 103 23  
## 7 1 5 46 151 12  
## 8 0 0 0 1 16

print(CF\_SVM)

## True\_Label  
## pred\_SVM 4 5 6 7 8  
## 4 3 4 0 0 0  
## 5 46 247 126 13 1  
## 6 23 163 500 177 35  
## 7 1 0 43 73 15  
## 8 0 0 0 0 0

print(CF\_DNN)

## True\_Label  
## pred\_DNN 4 5 6 7 8  
## 4 6 4 1 1 0  
## 5 45 273 160 17 1  
## 6 21 134 462 162 35  
## 7 1 2 45 83 14  
## 8 0 1 1 0 1

#metrics  
set.seed(2445)  
  
ml\_test\_KNN <- ml\_test(model\_KNN, True\_Label, output.as.table = FALSE)  
  
ml\_test\_NB <- ml\_test(pred\_NB, True\_Label, output.as.table = FALSE)  
  
ml\_test\_RF <- ml\_test(pred\_RF, True\_Label, output.as.table = FALSE)  
  
ml\_test\_SVM <- ml\_test(pred\_SVM, True\_Label, output.as.table = FALSE)  
  
ml\_test\_DNN <- ml\_test(pred\_DNN, True\_Label, output.as.table = FALSE)  
  
#Macro Average Accuracy  
  
MAvA\_KNN <- print((sum(ml\_test\_KNN$balanced.accuracy, na.rm = TRUE))/5)

## [1] 0.5100024

MAvA\_NB <- print((sum(ml\_test\_NB$balanced.accuracy, na.rm = TRUE))/5)

## [1] 0.5656745

MAvA\_RF <- print((sum(ml\_test\_RF$balanced.accuracy, na.rm = TRUE))/5)

## [1] 0.692838

MAvA\_SVM <- print((sum(ml\_test\_SVM$balanced.accuracy, na.rm = TRUE))/5)

## [1] 0.5788775

MAvA\_DNN <- print((sum(ml\_test\_DNN$balanced.accuracy, na.rm = TRUE))/5)

## [1] 0.5900534

#Macro Average F1   
  
MAvF1\_KNN <- print((sum(ml\_test\_KNN$F1, na.rm = TRUE))/5)

## [1] 0.2086379

MAvF1\_NB <- print((sum(ml\_test\_NB$F1, na.rm = TRUE))/5)

## [1] 0.3495931

MAvF1\_RF <- print((sum(ml\_test\_RF$F1, na.rm = TRUE))/5)

## [1] 0.5534347

MAvF1\_SVM <- print((sum(ml\_test\_SVM$F1, na.rm = TRUE))/5)

## [1] 0.3332035

MAvF1\_DNN <- print((sum(ml\_test\_DNN$F1, na.rm = TRUE))/5)

## [1] 0.3616275

#MAvMCC  
  
MAvMCC\_KNN <- print((sum(ml\_test\_KNN$MCC, na.rm = TRUE))/5)

## [1] 0.01992839

MAvMCC\_NB <- print((sum(ml\_test\_NB$MCC, na.rm = TRUE))/5)

## [1] 0.1430763

MAvMCC\_RF <- print((sum(ml\_test\_RF$MCC, na.rm = TRUE))/5)

## [1] 0.4646652

MAvMCC\_SVM <- print((sum(ml\_test\_SVM$MCC, na.rm = TRUE))/5)

## [1] 0.1867606

MAvMCC\_DNN <- print((sum(ml\_test\_DNN$MCC, na.rm = TRUE))/5)

## [1] 0.2215057

#MAvGeometricMean  
  
MAvGM\_KNN <- print((sum(ml\_test\_KNN$geometric.mean, na.rm = TRUE))/5)

## [1] 0.219316

MAvGM\_NB <- print((sum(ml\_test\_NB$geometric.mean, na.rm = TRUE))/5)

## [1] 0.4795046

MAvGM\_RF <- print((sum(ml\_test\_RF$geometric.mean, na.rm = TRUE))/5)

## [1] 0.632817

MAvGM\_SVM <- print((sum(ml\_test\_SVM$geometric.mean, na.rm = TRUE))/5)

## [1] 0.3919383

MAvGM\_DNN <- print((sum(ml\_test\_DNN$geometric.mean, na.rm = TRUE))/5)

## [1] 0.4485263