knn_lab.R

alexaubrey

Sun May 20 13:03:20 2018

```
library(class)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'zone/tz/2018c.
## 1.0/zoneinfo/America/Chicago'
```

```
data <- read.csv(file='balance.csv', head=FALSE, sep=",")

preProcess(data, method=c("center", "scale"))
```

```
## Created from 625 samples and 5 variables
##
## Pre-processing:
##   - centered (4)
##   - ignored (1)
##   - scaled (4)
```

```
t1 = sample(1:625, 500)
t2 = setdiff(1:625, t1)

c1 = data[t1,]$V1

train = subset(data[t1,], select =- V1)
test = subset(data[t2,], select =- V1)

pred <- knn(train, test, c1, k=3, prob=FALSE, use.all = TRUE)

confusionMatrix(pred, data[t2,]$V1)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  B  L  R
##          B  0  5  6
##          L  3 50  1
##          R  2  1 57
##
## Overall Statistics
##
##                Accuracy : 0.856
##                  95% CI : (0.782, 0.9124)
##     No Information Rate : 0.512
##     P-Value [Acc > NIR] : 5.565e-16
##
##                   Kappa : 0.7416
##  Mcnemar's Test P-Value : 0.4753
##
## Statistics by Class:
##
##                      Class: B Class: L Class: R
## Sensitivity            0.0000   0.8929   0.8906
## Specificity            0.9083   0.9420   0.9508
## Pos Pred Value         0.0000   0.9259   0.9500
## Neg Pred Value         0.9561   0.9155   0.8923
## Prevalence             0.0400   0.4480   0.5120
## Detection Rate         0.0000   0.4000   0.4560
## Detection Prevalence   0.0880   0.4320   0.4800
## Balanced Accuracy      0.4542   0.9174   0.9207
```

```
ktune <- train(train, c1, method="knn",
               tuneGrid=data.frame(.k = 1:20),
               trControl = trainControl(method = "cv"))

ktune
```

```
## k-Nearest Neighbors
##
## 500 samples
##   4 predictor
##   3 classes: 'B', 'L', 'R'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 451, 450, 449, 449, 451, 450, ...
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.8340936  0.7054652
##    2  0.8358928  0.7098298
##    3  0.8280936  0.6965216
##    4  0.8558207  0.7434590
##    5  0.8777815  0.7804396
##    6  0.8899024  0.7994028
##    7  0.9019432  0.8201200
##    8  0.9059448  0.8277389
##    9  0.9020640  0.8201313
##   10  0.9019840  0.8200174
##   11  0.9019840  0.8199533
##   12  0.9019840  0.8199885
##   13  0.8999840  0.8161108
##   14  0.8940232  0.8051610
##   15  0.8920232  0.8014139
##   16  0.8921032  0.8014804
##   17  0.8920624  0.8014922
##   18  0.8900624  0.7977552
##   19  0.8881032  0.7940635
##   20  0.8941441  0.8051911
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 8.
```

```
pred2 <- predict(ktune, test)
saveRDS(ktune, file="ktune.rds")

reload = readRDS("ktune.rds")
predict(reload, test)
```

```
##   [1] R R R R L R R R R R R R R L R R R R L L L L L R L L R R R R R R R R R
##  [36] B R R R R R R R R R L R R R R R R R R R L L L L L R R R L R R R R R R
##  [71] L R R L L L R L L L L L L L L L L L R L L R R R L L L L R L L L L L L R
## [106] L L L R R L L L L L L R L L L L L L L L
## Levels: B L R
```