**kNN-20237**

Jitesh Rawat

4/16/2021

**library**(palmerpenguins)  
**library**(tidyverse)

df <- penguins  
**head**(df)

## # A tibble: 6 x 8  
## species island bill\_length\_mm bill\_depth\_mm flipper\_length\_~ body\_mass\_g sex   
## <fct> <fct> <dbl> <dbl> <int> <int> <fct>  
## 1 Adelie Torge~ 39.1 18.7 181 3750 male   
## 2 Adelie Torge~ 39.5 17.4 186 3800 fema~  
## 3 Adelie Torge~ 40.3 18 195 3250 fema~  
## 4 Adelie Torge~ NA NA NA NA <NA>   
## 5 Adelie Torge~ 36.7 19.3 193 3450 fema~  
## 6 Adelie Torge~ 39.3 20.6 190 3650 male   
## # ... with 1 more variable: year <int>

df <- **na.omit**(df)  
**head**(df)

n = **nrow**(df)  
trainIndex = **sample**(1**:**n, size = **round**(0.8**\***n), replace=FALSE)  
train = df[trainIndex ,]  
test = df[**-**trainIndex ,]

**dim**(train)

## [1] 266 8

**dim**(test)

## [1] 67 8

**library**(caret)

*#?trainControl()*

Using cv  
trn\_ctrl <-**trainControl**(method = "cv", number = 10)  
  
model\_knn\_cv <- **train**(species**~**., data = train, method ="knn",  
 trControl = trn\_ctrl,   
 preProcess = **c**("center", "scale"),   
 tuneLength = 10)  
model\_knn\_cv

## k-Nearest Neighbors   
##   
## 266 samples  
## 7 predictor  
## 3 classes: 'Adelie', 'Chinstrap', 'Gentoo'   
##   
## Pre-processing: centered (8), scaled (8)   
## Resampling: Cross-Validated (10 fold)   
## Summary of sample sizes: 240, 240, 240, 240, 239, 239, ...   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 5 0.9813390 0.9700698  
## 7 0.9888889 0.9821192  
## 9 0.9888889 0.9821192  
## 11 0.9853175 0.9765747  
## 13 0.9853175 0.9765747  
## 15 0.9814713 0.9705700  
## 17 0.9817460 0.9711817  
## 19 0.9817460 0.9711817  
## 21 0.9778999 0.9657131  
## 23 0.9703500 0.9534763  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 9.

*#Validation on test data*  
prediction\_knn <- **predict**(model\_knn\_cv,newdata=test)  
**confusionMatrix**(prediction\_knn, reference = test**$**species, positive = "3")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Adelie Chinstrap Gentoo  
## Adelie 25 0 0  
## Chinstrap 0 14 0  
## Gentoo 0 0 28  
##   
## Overall Statistics  
##   
## Accuracy : 1   
## 95% CI : (0.9464, 1)  
## No Information Rate : 0.4179   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: Adelie Class: Chinstrap Class: Gentoo  
## Sensitivity 1.0000 1.000 1.0000  
## Specificity 1.0000 1.000 1.0000  
## Pos Pred Value 1.0000 1.000 1.0000  
## Neg Pred Value 1.0000 1.000 1.0000  
## Prevalence 0.3731 0.209 0.4179  
## Detection Rate 0.3731 0.209 0.4179  
## Detection Prevalence 0.3731 0.209 0.4179  
## Balanced Accuracy 1.0000 1.000 1.0000

*\**Using repeatedcv  
trn\_ctrl <-**trainControl**(method = "repeatedcv", number = 10, repeats = 3)  
  
model\_knn\_r\_cv <- **train**(species**~**., data = train, method ="knn",  
 trControl = trn\_ctrl,   
 preProcess = **c**("center", "scale"),   
 tuneLength = 10)  
model\_knn\_r\_cv

## k-Nearest Neighbors   
##   
## 266 samples  
## 7 predictor  
## 3 classes: 'Adelie', 'Chinstrap', 'Gentoo'   
##   
## Pre-processing: centered (8), scaled (8)   
## Resampling: Cross-Validated (10 fold, repeated 3 times)   
## Summary of sample sizes: 240, 240, 240, 239, 238, 239, ...   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 5 0.9898860 0.9839445  
## 7 0.9924027 0.9880309  
## 9 0.9911681 0.9860441  
## 11 0.9886515 0.9820276  
## 13 0.9886515 0.9820276  
## 15 0.9862705 0.9783987  
## 17 0.9861789 0.9782453  
## 19 0.9811932 0.9704989  
## 21 0.9774895 0.9650294  
## 23 0.9713133 0.9551513  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 7.

*#Validation on test data*   
prediction\_knn <- **predict**(model\_knn\_r\_cv,newdata=test)  
**confusionMatrix**(prediction\_knn, reference = test**$**species, positive = "3")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Adelie Chinstrap Gentoo  
## Adelie 25 0 0  
## Chinstrap 0 14 0  
## Gentoo 0 0 28  
##   
## Overall Statistics  
##   
## Accuracy : 1   
## 95% CI : (0.9464, 1)  
## No Information Rate : 0.4179   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: Adelie Class: Chinstrap Class: Gentoo  
## Sensitivity 1.0000 1.000 1.0000  
## Specificity 1.0000 1.000 1.0000  
## Pos Pred Value 1.0000 1.000 1.0000  
## Neg Pred Value 1.0000 1.000 1.0000  
## Prevalence 0.3731 0.209 0.4179  
## Detection Rate 0.3731 0.209 0.4179  
## Detection Prevalence 0.3731 0.209 0.4179  
## Balanced Accuracy 1.0000 1.000 1.0000

\*Using bootstrap

trn\_ctrl <-**trainControl**(method = "boot", number = 10)  
  
model\_knn\_boot <- **train**(species**~**., data = train, method ="knn",  
 trControl = trn\_ctrl,   
 preProcess = **c**("center", "scale"),   
 tuneLength = 10)  
model\_knn\_boot

## k-Nearest Neighbors   
##   
## 266 samples  
## 7 predictor  
## 3 classes: 'Adelie', 'Chinstrap', 'Gentoo'   
##   
## Pre-processing: centered (8), scaled (8)   
## Resampling: Bootstrapped (10 reps)   
## Summary of sample sizes: 266, 266, 266, 266, 266, 266, ...   
## Resampling results across tuning parameters:  
##   
## k Accuracy Kappa   
## 5 0.9827090 0.9726010  
## 7 0.9846900 0.9757095  
## 9 0.9846145 0.9754867  
## 11 0.9836436 0.9739681  
## 13 0.9825787 0.9722375  
## 15 0.9836645 0.9741333  
## 17 0.9826255 0.9725135  
## 19 0.9796241 0.9676915  
## 21 0.9755032 0.9611187  
## 23 0.9713976 0.9546730  
##   
## Accuracy was used to select the optimal model using the largest value.  
## The final value used for the model was k = 7.

*#Validation on test data*  
prediction\_knn <- **predict**(model\_knn\_boot,newdata=test)  
**confusionMatrix**(prediction\_knn, reference = test**$**species, positive = "3")

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction Adelie Chinstrap Gentoo  
## Adelie 25 0 0  
## Chinstrap 0 14 0  
## Gentoo 0 0 28  
##   
## Overall Statistics  
##   
## Accuracy : 1   
## 95% CI : (0.9464, 1)  
## No Information Rate : 0.4179   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: Adelie Class: Chinstrap Class: Gentoo  
## Sensitivity 1.0000 1.000 1.0000  
## Specificity 1.0000 1.000 1.0000  
## Pos Pred Value 1.0000 1.000 1.0000  
## Neg Pred Value 1.0000 1.000 1.0000  
## Prevalence 0.3731 0.209 0.4179  
## Detection Rate 0.3731 0.209 0.4179  
## Detection Prevalence 0.3731 0.209 0.4179  
## Balanced Accuracy 1.0000 1.000 1.0000

**plot**(model\_knn\_cv, main= "method = Cross Validation")



**plot**(model\_knn\_r\_cv, main= "method = RepeatedCV")



**plot**(model\_knn\_boot, main= "method = bootstrap")

