Cross Validation

Veerasak Kritsanapraphan

4/25/2018

Load Data

```
# load the libraries
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.3
```

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

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

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

```
library(klaR)
```

```
## Loading required package: MASS
```

```
# load the iris dataset
data(iris)
```

Split Data

Using Split Data

```
# define an 80%/20% train/test split of the dataset
split=0.70
trainIndex <- createDataPartition(iris$Species, p=split, list=FALSE)
data_train <- iris[ trainIndex,]
data_test <- iris[-trainIndex,]
# train a naive bayes model
model <- NaiveBayes(Species~., data=data_train)
# make predictions
x_test <- data_test[,1:4]
y_test <- data_test[,5]
predictions <- predict(model, x_test)
# summarize results
confusionMatrix(predictions$class, y_test)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         13         1
##   virginica       0          2        14
##
## Overall Statistics
##
```

```
##                Accuracy : 0.9333
##                  95% CI : (0.8173, 0.986)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.8667           0.9333
## Specificity                 1.0000            0.9667           0.9333
## Pos Pred Value              1.0000            0.9286           0.8750
## Neg Pred Value              1.0000            0.9355           0.9655
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.2889           0.3111
## Detection Prevalence        0.3333            0.3111           0.3556
## Balanced Accuracy           1.0000            0.9167           0.9333
```

Bootstraping

Using Bootstrap

```
# define training control
train_control <- trainControl(method="boot", number=10)
# train the model
model <- train(Species~., data=data_train,    trControl=train_control, method="nb")
# summarize results
print(model)
```

```
## Naive Bayes
##
## 105 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## No pre-processing
## Resampling: Bootstrapped (10 reps)
## Summary of sample sizes: 105, 105, 105, 105, 105, 105, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.9195750  0.8789905
##    TRUE      0.9288276  0.8930898
##
## Tuning parameter 'fL' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were fL = 0, usekernel = TRUE
##  and adjust = 1.
```

```r
predictions <- predict(model, x_test)
# summarize results
confusionMatrix(predictions, y_test)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         14         1
##   virginica       0          1        14
##
## Overall Statistics
##
##                Accuracy : 0.9556
##                  95% CI : (0.8485, 0.9946)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9333
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.9333           0.9333
## Specificity                 1.0000            0.9667           0.9667
## Pos Pred Value              1.0000            0.9333           0.9333
## Neg Pred Value              1.0000            0.9667           0.9667
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.3111           0.3111
## Detection Prevalence        0.3333            0.3333           0.3333
## Balanced Accuracy           1.0000            0.9500           0.9500
```

K-Fold Cross Validation

Using K-Fold Cross Validation

```r
# k-fold Cross Validation

# define training control
train_control <- trainControl(method="cv", number=10)
model <- train(Species~., data=data_train, trControl=train_control, method="nb")
# summarize results
print(model)
```

```
## Naive Bayes
##
## 105 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
```

```
## Summary of sample sizes: 95, 94, 93, 95, 94, 94, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.9438889  0.9164216
##    TRUE      0.9522222  0.9289216
##
## Tuning parameter 'fL' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were fL = 0, usekernel = TRUE
##  and adjust = 1.
```

```
predictions <- predict(model, x_test)
# summarize results
confusionMatrix(predictions, y_test)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         14         1
##   virginica       0          1        14
##
## Overall Statistics
##
##                Accuracy : 0.9556
##                  95% CI : (0.8485, 0.9946)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9333
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.9333           0.9333
## Specificity                 1.0000            0.9667           0.9667
## Pos Pred Value              1.0000            0.9333           0.9333
## Neg Pred Value              1.0000            0.9667           0.9667
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.3111           0.3111
## Detection Prevalence        0.3333            0.3333           0.3333
## Balanced Accuracy           1.0000            0.9500           0.9500
```

Repeated K-Fold Cross Validation

Using Repeated K-Fold Cross Validation

```
# define training control
train_control <- trainControl(method="repeatedcv", number=10, repeats=3)
```

```r
# train the model
model <- train(Species~., data=data_train, trControl=train_control, method="nb")
# summarize results
print(model)
```

```
## Naive Bayes
##
## 105 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold, repeated 3 times)
## Summary of sample sizes: 95, 94, 95, 94, 93, 95, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy   Kappa
##   FALSE      0.9569865  0.9351360
##    TRUE      0.9543266  0.9311143
##
## Tuning parameter 'fL' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were fL = 0, usekernel = FALSE
##  and adjust = 1.
```

```r
predictions <- predict(model, x_test)
# summarize results
confusionMatrix(predictions, y_test)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         13         1
##   virginica       0          2        14
##
## Overall Statistics
##
##                Accuracy : 0.9333
##                  95% CI : (0.8173, 0.986)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.8667           0.9333
## Specificity                 1.0000            0.9667           0.9333
```

```
## Pos Pred Value              1.0000            0.9286           0.8750
## Neg Pred Value              1.0000            0.9355           0.9655
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.2889           0.3111
## Detection Prevalence        0.3333            0.3111           0.3556
## Balanced Accuracy           1.0000            0.9167           0.9333
```

Leave one out Cross Validation

Using Leave one out Cross Validation

```
# define training control
train_control <- trainControl(method="LOOCV")
# train the model
model <- train(Species~., data=data_train, trControl=train_control, method="nb")
# summarize results
print(model)
```

```
## Naive Bayes
##
## 105 samples
##   4 predictor
##   3 classes: 'setosa', 'versicolor', 'virginica'
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 104, 104, 104, 104, 104, 104, ...
## Resampling results across tuning parameters:
##
##   usekernel  Accuracy  Kappa
##   FALSE      0.952381  0.9285714
##    TRUE      0.952381  0.9285714
##
## Tuning parameter 'fL' was held constant at a value of 0
## Tuning
##  parameter 'adjust' was held constant at a value of 1
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were fL = 0, usekernel = FALSE
##  and adjust = 1.
```

```
predictions <- predict(model, x_test)
# summarize results
confusionMatrix(predictions, y_test)
```

```
## Confusion Matrix and Statistics
##
##             Reference
## Prediction   setosa versicolor virginica
##   setosa         15          0         0
##   versicolor      0         13         1
##   virginica       0          2        14
##
## Overall Statistics
##
##                Accuracy : 0.9333
```

```
##                  95% CI : (0.8173, 0.986)
##     No Information Rate : 0.3333
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: setosa Class: versicolor Class: virginica
## Sensitivity                 1.0000            0.8667           0.9333
## Specificity                 1.0000            0.9667           0.9333
## Pos Pred Value              1.0000            0.9286           0.8750
## Neg Pred Value              1.0000            0.9355           0.9655
## Prevalence                  0.3333            0.3333           0.3333
## Detection Rate              0.3333            0.2889           0.3111
## Detection Prevalence        0.3333            0.3111           0.3556
## Balanced Accuracy           1.0000            0.9167           0.9333
```