R Notebook

# Testes de Kernel do SVM

Neste documento são feitos testes dos diferentes kernels presentes na função svm do pacote e1071 para a construção de suport vector machines. A base de dados usada foi a Custumer Churn. Tirando o customerID, todos os atributos foram usados para classificar os indivíduos em relação ao campo Churn.

## Preparação e separação do data set em subsets de treinamento e teste.

Foi retirada as linhas em que o campo TotalCharges se encontrava como NA, ou seja, não haviam informações sobre o atributo. O dataset foi dividido em um subconjunto com 80% dos dados totais para treinamento, e um subconjunto de 20% dos dados para teste.

require(e1071)

## Loading required package: e1071

require(caret)

## Loading required package: caret

## Loading required package: lattice

## Loading required package: ggplot2

require(dplyr)

## Loading required package: dplyr

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

dados <- read.csv("../dados/dados.csv")  
dados <- tbl\_df(dados)  
dados$customerID <- as.character(dados$customerID)  
  
DataSubset <- subset(dados,subset=TotalCharges>0,select = -customerID)  
DadosTreinamento <- createDataPartition(y = DataSubset$Churn,p = 0.8)  
Train <- DataSubset[DadosTreinamento$Resample1,]  
Test <- DataSubset[-DadosTreinamento$Resample1,]

## Kernel sigmoid do SVM

preditor <- svm(Churn ~ ., data = Train,kernel="sigmoid")  
preditions <- predict(preditor,subset(Test,select = -Churn))  
confusionMatrix(preditions,subset(Test,select=Churn)$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 898 148  
## Yes 134 225  
##   
## Accuracy : 0.7993   
## 95% CI : (0.7774, 0.8199)  
## No Information Rate : 0.7345   
## P-Value [Acc > NIR] : 9.27e-09   
##   
## Kappa : 0.4791   
## Mcnemar's Test P-Value : 0.4388   
##   
## Sensitivity : 0.8702   
## Specificity : 0.6032   
## Pos Pred Value : 0.8585   
## Neg Pred Value : 0.6267   
## Prevalence : 0.7345   
## Detection Rate : 0.6391   
## Detection Prevalence : 0.7445   
## Balanced Accuracy : 0.7367   
##   
## 'Positive' Class : No   
##

## Kernel linear do SVM

preditor <- svm(Churn ~ ., data = Train,kernel="linear")  
preditions <- predict(preditor,subset(Test,select = -Churn))  
confusionMatrix(preditions,subset(Test,select=Churn)$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 913 152  
## Yes 119 221  
##   
## Accuracy : 0.8071   
## 95% CI : (0.7855, 0.8274)  
## No Information Rate : 0.7345   
## P-Value [Acc > NIR] : 1.183e-10   
##   
## Kappa : 0.4911   
## Mcnemar's Test P-Value : 0.05191   
##   
## Sensitivity : 0.8847   
## Specificity : 0.5925   
## Pos Pred Value : 0.8573   
## Neg Pred Value : 0.6500   
## Prevalence : 0.7345   
## Detection Rate : 0.6498   
## Detection Prevalence : 0.7580   
## Balanced Accuracy : 0.7386   
##   
## 'Positive' Class : No   
##

## Kernel polynomial do SVM

preditor <- svm(Churn ~ ., data = Train,kernel="polynomial")  
preditions <- predict(preditor,subset(Test,select = -Churn))  
confusionMatrix(preditions,subset(Test,select=Churn)$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 975 230  
## Yes 57 143  
##   
## Accuracy : 0.7957   
## 95% CI : (0.7737, 0.8165)  
## No Information Rate : 0.7345   
## P-Value [Acc > NIR] : 5.654e-08   
##   
## Kappa : 0.3852   
## Mcnemar's Test P-Value : < 2.2e-16   
##   
## Sensitivity : 0.9448   
## Specificity : 0.3834   
## Pos Pred Value : 0.8091   
## Neg Pred Value : 0.7150   
## Prevalence : 0.7345   
## Detection Rate : 0.6940   
## Detection Prevalence : 0.8577   
## Balanced Accuracy : 0.6641   
##   
## 'Positive' Class : No   
##

## Kernel radial basis do SVM

preditor <- svm(Churn ~ ., data = Train,kernel="radial")  
preditions <- predict(preditor,subset(Test,select = -Churn))  
confusionMatrix(preditions,subset(Test,select=Churn)$Churn)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 940 184  
## Yes 92 189  
##   
## Accuracy : 0.8036   
## 95% CI : (0.7818, 0.824)  
## No Information Rate : 0.7345   
## P-Value [Acc > NIR] : 9.176e-10   
##   
## Kappa : 0.4532   
## Mcnemar's Test P-Value : 4.312e-08   
##   
## Sensitivity : 0.9109   
## Specificity : 0.5067   
## Pos Pred Value : 0.8363   
## Neg Pred Value : 0.6726   
## Prevalence : 0.7345   
## Detection Rate : 0.6690   
## Detection Prevalence : 0.8000   
## Balanced Accuracy : 0.7088   
##   
## 'Positive' Class : No   
##

## Resultados

Em relação ao atributo kappa de cada kernel:

Sigmoid: 0.4459

Linear: 0.4662

Polynomial: 0.3882

Radial basis: 0.4501

Em relação a este atributo, o kernel Linear é a melhor opção para uma SVM.