SVM for Spambase data

# 1. Spambase 데이터

# 데이터 읽고 분할   
setwd("D:/Dropbox/PDSwR/Spambase/")  
spamD <- read.table('spamD.tsv', header=T, sep='\t')  
spamTrain <- subset(spamD, spamD$rgroup>=10)  
spamTest <- subset(spamD, spamD$rgroup<10)  
  
## 로지스틱 회  
spamVars <- setdiff(colnames(spamD),list('rgroup','spam'))  
spamFormula <- as.formula(paste('spam=="spam"',  
 paste(spamVars,collapse=' + '),sep=' ~ '))  
  
spamModel <- glm(spamFormula,family=binomial(link='logit'),  
 data=spamTrain)  
# 시험데이터   
spamTest$pred <- predict(spamModel,newdata=spamTest,  
 type='response')  
print(with(spamTest,table(y=spam,glPred=pred>=0.5)))

## glPred  
## y FALSE TRUE  
## non-spam 264 14  
## spam 22 158

## 지지벡터기계  
library(kernlab)  
spamFormulaV <- as.formula(paste('spam',  
 paste(spamVars,collapse=' + '),sep=' ~ '))  
  
svmM <- ksvm(spamFormulaV,data=spamTrain,   
 kernel='rbfdot', # 가우스 커널  
 C=10,   
 prob.model=T, cross=5,   
 class.weights=c('spam'=1,'non-spam'=10) # 가중치   
 )  
spamTest$svmPred <- predict(svmM,newdata=spamTest,type='response')  
print(with(spamTest,table(y=spam,svmPred=svmPred)))

## svmPred  
## y non-spam spam  
## non-spam 269 9  
## spam 27 153

print(svmM)

## Support Vector Machine object of class "ksvm"   
##   
## SV type: C-svc (classification)   
## parameter : cost C = 10   
##   
## Gaussian Radial Basis kernel function.   
## Hyperparameter : sigma = 0.030136090137948   
##   
## Number of Support Vectors : 1119   
##   
## Objective Function Value : -4626.691   
## Training error : 0.02824   
## Cross validation error : 0.078931   
## Probability model included.

# 2. Spiral 예제

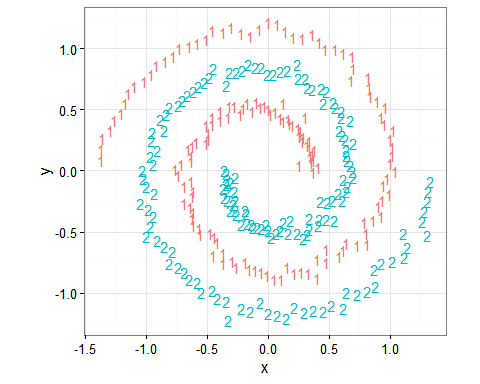
kernlab 패키지의 예제

library(kernlab)  
data(spirals)   
  
sc <- specc(spirals, centers = 2)   
s <- data.frame(x=spirals[,1],y=spirals[,2],  
 class=as.factor(sc))   
  
# 예  
library('ggplot2')

##   
## Attaching package: 'ggplot2'

## The following object is masked from 'package:kernlab':  
##   
## alpha

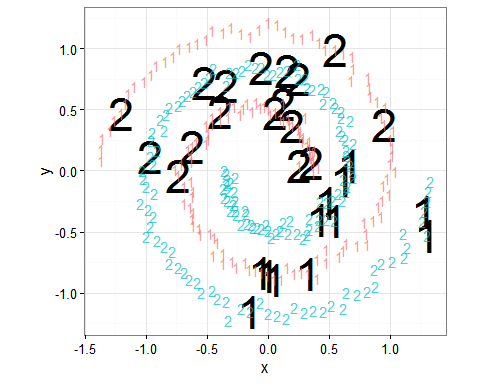
ggplot(data=s) +  
 geom\_text(aes(x=x,y=y,  
 label=class,color=class)) +  
 coord\_fixed() +   
 theme\_bw() + theme(legend.position='none')



# 데이터 생성 및 분할  
set.seed(2335246L)  
s$group <- sample.int(100,size=dim(s)[[1]],replace=T)  
sTrain <- subset(s,group>10)  
sTest <- subset(s,group<=10)

선형 커널

library(e1071)  
mSVMV <- svm(class~x+y,data=sTrain,kernel='linear',type='nu-classification')   
sTest$predSVMV <- predict(mSVMV,newdata=sTest,type='response')   
  
ggplot() +  
 geom\_text(data=sTest,aes(x=x,y=y,  
 label=predSVMV),size=12) +  
 geom\_text(data=s,aes(x=x,y=y,  
 label=class,color=class),alpha=0.7) +  
 coord\_fixed() +   
 theme\_bw() + theme(legend.position='none')



## 연습문제

1. Spiral 예제에 대하여 가우스 커널 지지벡터기계를 적합하고 선형 커널의 결과와 비교하시오.
2. GAM 로지스틱 회귀를 적합하고 지지벡터기계와 비교하시오.