Model Selection

Kanchana Jagannathan

1/17/2019

R Markdown

```
library(caret)
```

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

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

```
mydata <- read.csv('winequality-red.csv',sep=';')

mydata$quality <- as.factor(mydata$quality)
control <- trainControl(method = "repeatedcv",number = 10, repeats = 3)
seed <- 123

metric <- "Accuracy"
```

Split the data into training and test set

```
training.samples <- createDataPartition(mydata$quality, p = 0.8, list = FALSE)
train.data  <- mydata[training.samples, ]
test.data <- mydata[-training.samples, ]
```

Model Building

```
pre_process = c("center","scale")

# Linear Discriminant Analysis
set.seed(seed)
fit.lda<-train(quality~., data=train.data,method="lda",metric=metric,
               preProc=c("center","scale"),trControl=control)

# SVM Radial
set.seed(seed)
fit.svmRadial<-train(quality~., data=train.data,method="svmRadial",metric=metric,
               preProc=c("center","scale"),trControl=control,fit=FALSE)

# knn
set.seed(seed)
fit.knn<-train(quality~., data=train.data,method="knn",metric=metric,
               preProc=c("center","scale"),trControl=control)

# CART
set.seed(seed)
```

```
fit.cart<-train(quality~., data=train.data,method="rpart",metric=metric,
                trControl=control)

# C5.0
set.seed(seed)
fit.c50<-train(quality~., data=train.data,method="C5.0",metric=metric,
                trControl=control)

# Bagged CART
set.seed(seed)
fit.treebag<-train(quality~., data=train.data,method="treebag",metric=metric,
                trControl=control)
# Random Forest
set.seed(seed)
fit.rf<-train(quality~., data=train.data,method="rf",metric=metric,
                trControl=control)

#Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(quality~., data=train.data, method="gbm", metric=metric, trControl=control, verbose=FA
```

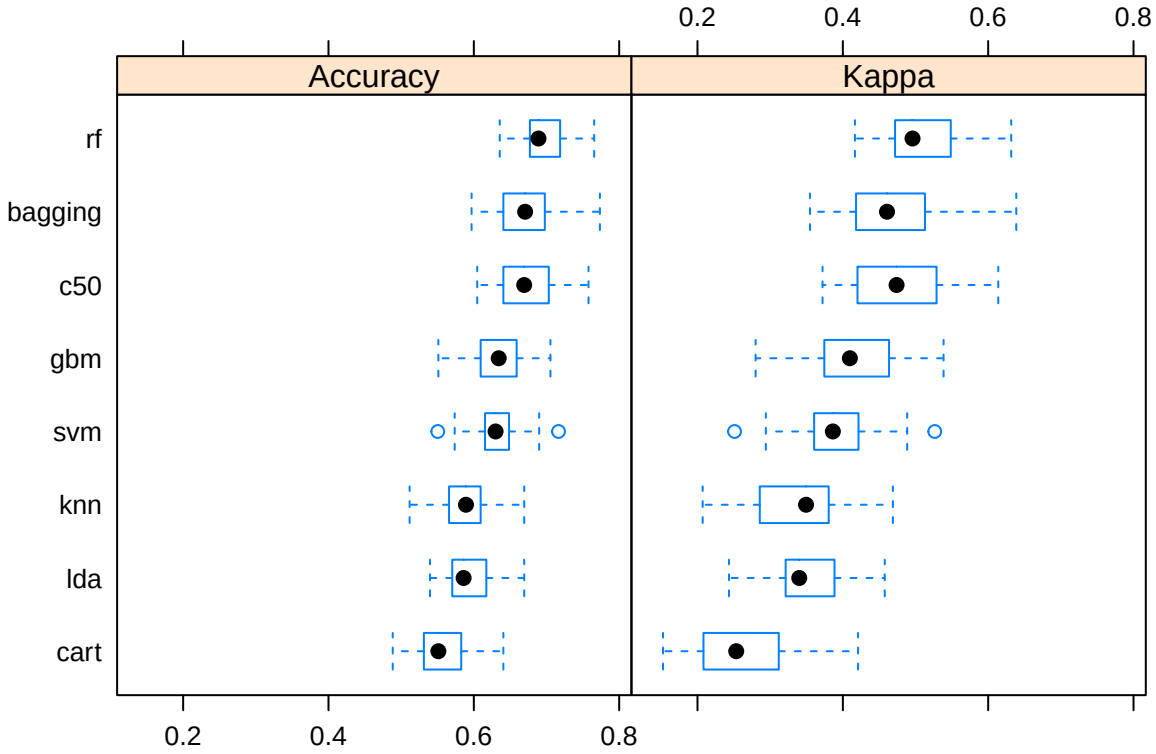
results

```
results <- resamples(list(lda=fit.lda, svm=fit.svmRadial, knn=fit.knn,
                          cart=fit.cart, c50=fit.c50,
                          bagging=fit.treebag, rf=fit.rf, gbm=fit.gbm))
# Table comparison
print(summary(results))
```

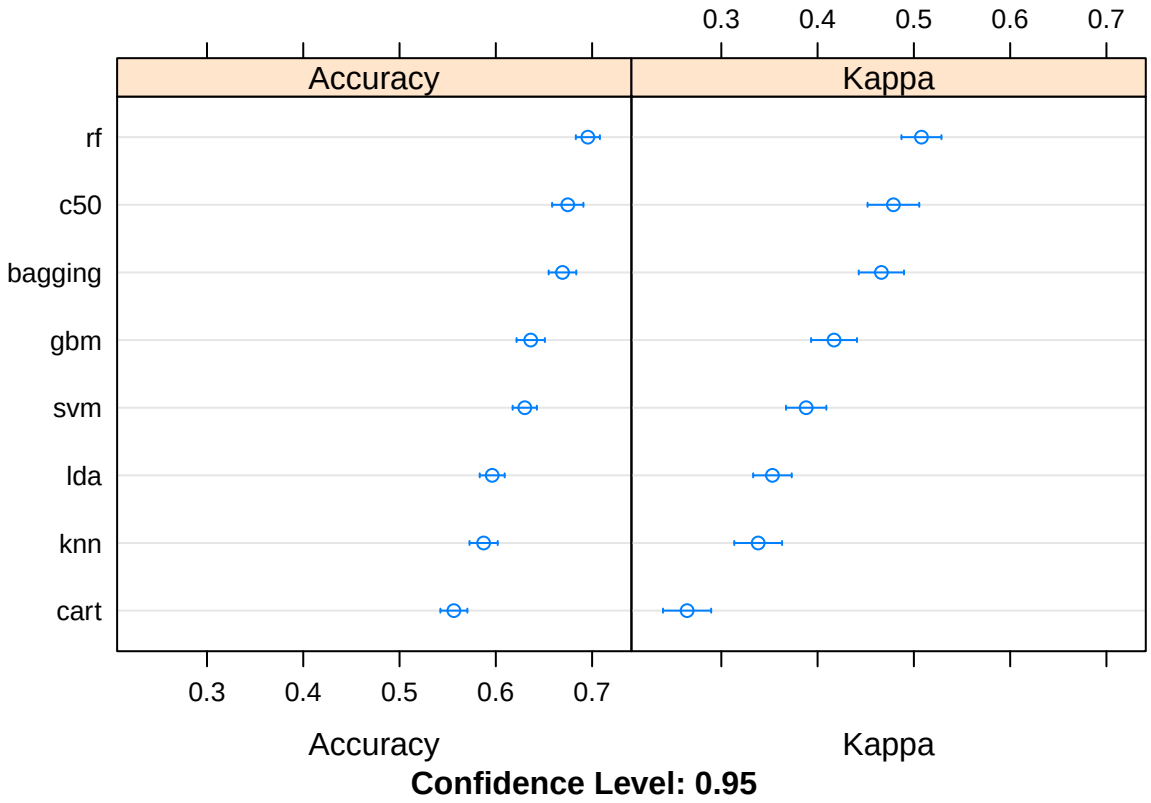
```
##
## Call:
## summary.resamples(object = results)
##
## Models: lda, svm, knn, cart, c50, bagging, rf, gbm
## Number of resamples: 30
##
## Accuracy
##              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda     0.5396825 0.5703125 0.5859375 0.5962278 0.6171875 0.6692913    0
## svm     0.5503876 0.6158353 0.6299213 0.6300539 0.6477454 0.7165354    0
## knn     0.5116279 0.5661509 0.5891473 0.5873108 0.6074219 0.6692913    0
## cart    0.4883721 0.5312500 0.5511811 0.5564273 0.5812386 0.6406250    0
## c50     0.6046512 0.6418861 0.6692913 0.6747170 0.7017624 0.7578125    0
## bagging 0.5968992 0.6413215 0.6705832 0.6692260 0.6970839 0.7734375    0
## rf      0.6356589 0.6771654 0.6889261 0.6955742 0.7162279 0.7656250    0
## gbm     0.5511811 0.6093750 0.6342660 0.6362347 0.6589147 0.7054264    0
##
## Kappa
##              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## lda     0.2433216 0.3216740 0.3400193 0.3530896 0.3843508 0.4577064    0
## svm     0.2509761 0.3634638 0.3862651 0.3881186 0.4183719 0.5264630    0
```

```
## knn      0.2071136 0.2883309 0.3493779 0.3382697 0.3775537 0.4688817    0
## cart     0.1525829 0.2133177 0.2533058 0.2644592 0.3081178 0.4209284    0
## c50      0.3720531 0.4220284 0.4739211 0.4787235 0.5259189 0.6138682    0
## bagging 0.3547737 0.4192067 0.4607875 0.4662547 0.5122552 0.6386645    0
## rf       0.4166339 0.4720306 0.4959170 0.5078534 0.5436855 0.6316547    0
## gbm      0.2799881 0.3750586 0.4096553 0.4170702 0.4630927 0.5386075    0
```

```r
# boxplot comparison
bwplot(results)
```




```r
# Dot-plot comparison
dotplot(results)
```



Confidence Level: 0.95