RWekaJ48.R

Administrator

2019-11-06

library(RWeka)  
trainset <- read.csv("C:/Users/Administrator/Documents/MSDatascience/IST707/Week5/titanic-train.csv")  
testset <- read.csv("C:/Users/Administrator/Documents/MSDatascience/IST707/Week5/titanic-test.csv")  
NN <- make\_Weka\_filter("weka/filters/unsupervised/attribute/NumericToNominal")  
trainset <- NN(data=trainset, control= Weka\_control(R="1-3"), na.action = NULL)  
testset <- NN(data=testset, control= Weka\_control(R="1,3"), na.action = NULL)  
MS <- make\_Weka\_filter("weka/filters/unsupervised/attribute/ReplaceMissingValues")   
trainset <-MS(data=trainset, na.action = NULL)  
testset <-MS(data=testset, na.action = NULL)  
  
#model 1: no feature engineering  
  
m=J48(Survived~., data = trainset, control=Weka\_control(U=FALSE, M=2, C=0.5))  
e <- evaluate\_Weka\_classifier(m, numFolds = 10, seed = 1, class = TRUE)  
pred=predict (m, newdata = testset, type = c("class"))  
myids=c("PassengerId")  
id\_col=testset[myids]  
newpred=cbind(id\_col, pred)  
colnames(newpred)=c("Passengerid", "Survived")  
write.csv(newpred, file="C:/Users/Administrator/Documents/MSDatascience/IST707/Week5/titanic-J48-pred-11-06-2019.csv", row.names=FALSE)  
  
#80.8%  
#Kaggle .78469  
  
#model 2: some feature selection  
  
myVars=c("Pclass", "Sex", "Age", "SibSp", "Parch", "Fare", "Survived")  
newtrain=trainset[myVars]  
newtest=testset[myVars]  
m=J48(Survived~., data = newtrain)  
m=J48(Survived~., data = newtrain, control=Weka\_control(U=FALSE, M=2, C=0.5))  
e=evaluate\_Weka\_classifier(m, seed=1, numFolds=10)  
pred=predict (m, newdata = newtest, type = c("class"))  
myids=c("PassengerId")  
id\_col=testset[myids]  
newpred=cbind(id\_col, pred)  
colnames(newpred)=c("Passengerid", "Survived")  
View(newpred)  
write.csv(newpred, file="C:/Users/Administrator/Documents/MSDatascience/IST707/Week5/titanic-J48-pred-11-06-2019-fs.csv", row.names=FALSE)  
  
InfoGainAttributeEval(Survived ~ . , data= trainset)

## Pclass Sex Age SibSp Parch Ticket   
## 0.08383105 0.21766011 0.01670457 0.02650997 0.01538075 0.82593692   
## Fare Cabin Embarked PassengerId   
## 0.09616054 0.23838554 0.02404709 0.96070790

#81.7%  
#Kaggle .78469