Tutorial: naive Bayes in Pakcage e1071 for Titanic Prediction

Bei Yu

02/14/2017

This is a tutorial on using the naive Bayes algorithm in the e1071 package to predict Titanic survivors.

prepare data

First load the training data in csv format, and then convert "Survived" to nominal variable and "Pclass" to ordinal variable.

```
trainset <- read.csv("/Users/byu/Desktop/Data/titanic-train.csv")
trainset$Survived=factor(trainset$Survived)
trainset$Pclass=ordered(trainset$Pclass)
```

Then load the test data and convert attributes in similar way.

```
testset <- read.csv("/Users/byu/Desktop/Data/titanic-test.csv")
testset$Survived=factor(testset$Survived)
testset$Pclass=ordered(testset$Pclass)
```

Then remove some attributes that are not likely to be helpful, such as "embarked" - create a new data set with all other attributes. Process the train and test set in the same way.

```
myVars=c("Pclass", "Sex", "Age", "SibSp", "Fare", "Survived")
newtrain=trainset[myVars]
newtest=testset[myVars]
```

naive Bayes in e1071

Now load the package e1071

```
library(e1071)
```

Build naive Bayes model using the e1071 package

```
nb=naiveBayes(Survived~., data = newtrain, laplace = 1, na.action = na.pass)
```

Apply the model to predicting test data

```
pred=predict(nb, newdata=newtest, type=c("class"))
```

Combine the predictions with the corresponding case ids.

```
myids=c("PassengerId")
id_col=testset[myids]
newpred=cbind(id_col, pred)
```

Add header to output

```
colnames(newpred)=c("Passengerid", "Survived")
```

Write output to file

```
write.csv(newpred, file="/Users/byu/Desktop/Data/titanic-NB-pred.csv",
row.names=FALSE)
```

For more information about naive Bayes in e1071, see the manual at https://cran.r-project.org/web/packages/e1071/e1071.pdf