Chapter 8 – Naïve Bayes

Instructor: Zach Zhizhong ZHOU,
Shanghai Jiao Tong University
主讲教师：周志中，上海交通大学

Data Mining for Business Intelligence

Shmueli, Patel & Bruce

FlightDelays - 1

```
library(e1071) ## needed for Naive Bayes
library(lattice) ## needed for package caret
library(ggplot2) ## needed for package caret
library(caret) ## needed for tuning Naive Bayes models
library(combinat) ## needed for package klaR
library(MASS) ## needed for package klaR
library(klaR) ## needed for tuning Naive Bayes models
setwd("C:/BA/NaiveBayes")
FlightDelays <- read.csv("FlightDelays.csv",header = TRUE)
names(FlightDelays)
```

FlightDelays - 2

```
FlightDelays[,"Weather"] <- factor(FlightDelays[,"Weather"])

FlightDelays[,"DAY_WEEK"] <-
factor(FlightDelays[,"DAY_WEEK"])



SchDepHour <- floor(FlightDelays[,"CRS_DEP_TIME"]/100)

SchDepHour <- factor(SchDepHour)

FlightDelays <- cbind(FlightDelays,SchDepHour)



summary(FlightDelays)

sapply(FlightDelays,class)
```

FlightDelays - 3

```
Predictors <-
FlightDelays[,c("ORIGIN","DEST","CARRIER","Weather",

                        "DAY_WEEK","SchDepHour")]



classifier<-naiveBayes(Predictors,
FlightDelays[,"Flight.Status"])

table(predict(classifier, Predictors),
FlightDelays[,"Flight.Status"])
```

FlightDelays - 4

model <- train(Predictors, FlightDelays[,"Flight.Status"],
#先列出Predictors,然后是结果变量，此处是Flight.Status。
method='nb',# method = 'nb' 指的是使用Naive Bayes
metric='Accuracy', #评价指标是“准确率” Accuracy
trControl=trainControl(method='repeatedcv',
number=10, repeats=2))
#trControl 是对训练过程进行控制的函数。此处的
method='repeatedcv'意思是使用repeated cross validation
#方法（重复交叉验证）。number=10表示做10-fold cross
validation，意思是把数据集割成10块，然后做10次
#训练和验证，每次都取其中一块数据（1/10的数据）当验证数
据集，剩下的当训练数据集。repeat=2表示上面的
#过程重复2次，等总共要做20次训练-验证。最终计算评价指标
（此处是Accuracy）的平均值。

FlightDelays - 5

```
model

confusionMatrix(model)



Predictors_Example <- data.frame(ORIGIN="DCA",
DEST="JFK", CARRIER = "US",  Weather = "1",
DAY_WEEK = "3", SchDepHour = "17")

predict(model$finalModel,Predictors_Example)
```