# Naive Bayes on the Titanic Data

*Karen Mazidi*

Performing Naive Bayes on the Titanic data set.

### Load and preprocess the data

We will skip the description of these steps since they are the same as in the logistic regression example.

```
df <- read.csv("data/titanic3.csv", header=TRUE)

# data cleaning
df <- df[,c(1,2,4,5)]
df$pclass <- factor(df$pclass)
df$survived <- factor(df$survived)

# handle missing values
df <- df[!is.na(df$pclass),]
df <- df[!is.na(df$survived),]
df$age[is.na(df$age)] <- median(df$age,na.rm=T)
```

### Divide into train and test sets

This should be the same split as we had for logistic regression so we can compare the two algorithms.

```
set.seed(1234)
i <- sample(1:nrow(df), 0.75*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

### Build the model

The naive Bayes algorithm is in package e1071.

```
library(e1071)
nb1 <- naiveBayes(survived~., data=train)
nb1
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##        0        1
## 0.617737 0.382263
##
## Conditional probabilities:
##    pclass
## Y           1         2         3
##   0 0.1468647 0.1930693 0.6600660
##   1 0.3946667 0.2426667 0.3626667
```

```
##
##    sex
## Y                female      male
##   0 0.0000000 0.1584158 0.8415842
##   1 0.0000000 0.6773333 0.3226667
##
##    age
## Y       [,1]     [,2]
##   0 30.32109 12.32909
##   1 28.14467 13.83251
```

### Evaluate on the test data

```
p1 <- predict(nb1, newdata=test, type="class")
table(p1, test$survived)
```

```
##
## p1    0   1
##   0 172  39
##   1  31  86
```

```
mean(p1==test$survived)
```

```
## [1] 0.7865854
```

We got very slightly higher for naive Bayes than we did for logistic regression.

### Extracting probabilities

One of the nice things about the algorithm is that you can extract the raw probabilities.

```
p1_raw <- predict(nb1, newdata=test, type="raw")
head(p1_raw)
```

```
##               0         1
## [1,] 0.06305836 0.9369416
## [2,] 0.12856023 0.8714398
## [3,] 0.65142708 0.3485729
## [4,] 0.12742936 0.8725706
## [5,] 0.11136485 0.8886352
## [6,] 0.37141738 0.6285826
```

### Remove Age

When we look at the Naive Bayes algorithm, we see the mean for survived versus perished is different by only one year. This suggests that age has little predictive value. Let's check that by building another model, this time without age.

```
nb2 <- naiveBayes(survived~.-age, data=train)
p2 <- predict(nb2, newdata=test[,-4], type="class")
table(p2, test$survived)
```

```
##
## p2    0   1
##   0 172  40
##   1  31  85
```

```r
mean(p2==test$survived)
```

```
## [1] 0.7835366
```

As it turns out, there is only a very slight difference in the accuracies: .8079 to .8109, both round to 81%.