secondAttempt

isilendil

2014.08.16

load train set

```
train <- read.csv("~/Kaggle/Titanic:MachineLearningfromDisaster/titanic/train.csv")
```

load test set

```
test <- read.csv("~/Kaggle/Titanic:MachineLearningfromDisaster/titanic/test.csv")
```

summary results

```
table(train$Sex)
```

```
##
## female   male
##    314    577
```

```
prop.table(table(train$Sex))
```

```
##
## female   male
## 0.3524 0.6476
```

```
table(train$Sex, train$Survived)
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

```
prop.table(table(train$Sex, train$Survived))
```

```
##
##                0       1
##   female 0.09091 0.26150
##   male   0.52525 0.12233
```

proportions in the 1st dimension

```r
table(train$Sex, train$Survived)
```

```
##
##            0   1
##   female  81 233
##   male   468 109
```

```r
prop.table(table(train$Sex, train$Survived), 1)
```

```
##
##                0      1
##   female 0.2580 0.7420
##   male   0.8111 0.1889
```

predict according to the gender

test$Survived < -0test$Survived[test$Sex == "female"] <- 1

summary of ages

```r
summary(train$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.42   20.10   28.00   29.70   38.00   80.00     177
```

create a new variable to indicate whether the passenger is a child(< 18)

```r
train$Child <- 0
train$Child[train$Age < 18] <- 1
```

the number of survivors in different groups

```r
aggregate(Survived ~ Child + Sex, data = train, FUN = sum)
```

```
##   Child    Sex Survived
## 1     0 female      195
## 2     1 female       38
## 3     0   male       86
## 4     1   male       23
```

the number of passengers in different groups

```r
aggregate(Survived ~ Child + Sex, data = train, FUN = length)
```

```
##   Child    Sex Survived
## 1     0 female      259
## 2     1 female       55
## 3     0   male      519
## 4     1   male       58
```

the proportion of survivors in different groups

```r
aggregate(Survived ~ Child + Sex, data = train, FUN = function(x) { sum(x) / length(x) })
```

```
##   Child    Sex Survived
## 1     0 female   0.7529
## 2     1 female   0.6909
## 3     0   male   0.1657
## 4     1   male   0.3966
```

summary of the fare of tickets

```r
summary(train$Fare)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.0     7.9    14.5    32.2    31.0   512.0
```

create a new variable to hash the fare to different intervals

```r
train$Fare2 <- "30+"
train$Fare2[train$Fare < 30 & train$Fare >= 20] <- "20-30"
train$Fare2[train$Fare < 20 & train$Fare >= 10] <- "10-20"
train$Fare2[train$Fare < 10] <- "<10"
```

aggregate result about Survived related to Fare2, Sex and Pclass

```r
aggregate(Survived ~ Fare2 + Pclass + Sex, data = train, FUN = function(x) { sum(x) / length(x) })
```

```
##    Fare2 Pclass    Sex Survived
## 1  20-30      1 female   0.8333
## 2    30+      1 female   0.9773
## 3  10-20      2 female   0.9143
## 4  20-30      2 female   0.9000
## 5    30+      2 female   1.0000
## 6    <10      3 female   0.5938
## 7  10-20      3 female   0.5814
```

```
## 8  20-30      3 female   0.3333
## 9    30+      3 female   0.1250
## 10   <10      1   male   0.0000
## 11 20-30      1   male   0.4000
## 12   30+      1   male   0.3837
## 13   <10      2   male   0.0000
## 14 10-20      2   male   0.1587
## 15 20-30      2   male   0.1600
## 16   30+      2   male   0.2143
## 17   <10      3   male   0.1115
## 18 10-20      3   male   0.2368
## 19 20-30      3   male   0.1250
## 20   30+      3   male   0.2400
```

a new prediction

```
test$Survived <- 0
test$Survived[test$Sex == "female"] <- 1
test$Survived[test$Sex == "female" & test$Pclass == 3 & test$Fare >= 20] <- 0
```

generate result

```
submit <- data.frame(PassengerId = test$PassengerId, Survived = test$Survived)
write.csv(submit, file = "/home/yanyg/Kaggle/Titanic:MachineLearningfromDisaster/titanic/secondAttempt.
```