Assignment_3

Rupesh_Suragani

2023-10-16

Summary

Questions - Answers

1. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why? A. From the above data we can conclude that the accidents that are Injured is 21462 i.e. 50.88%.

---

2. Select the first 24 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER_R and TRAF_CON_R. Create a pivot table that examines INJURY as a function of the two predictors for these 24 records. Use all three variables in the pivot table as rows/columns.

2.1. Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors. A. Possible combinations of the predictors are Probability injury=yes when weather=1, traffic=0 is 0.6666667 Probability injury=yes when weather=2, traffic=0 is 0.1818182 Probability injury=yes when weather=1, traffic=1 is 0 Probability injury=yes when weather=2, traffic=1 is 0 Probability injury=yes when weather=1, traffic=2 is 0 Probability injury=yes when weather=2, traffic=2 is 1

2.2. Classify the 24 accidents using these probabilities and a cutoff of 0.5. A. Out of 24 rows there are 5 rows that the actual values of Injury from the dataset that does not matches with the predicted values. A. Quantitative predictions- 0.6666667 0.1818182 0.0000000 0.0000000 0.6666667 0.1818182 0.1818182 0.6666667 0.1818182 0.1818182 0.1818182 0.0000000 0.6666667 0.6666667 0.6666667 0.6666667 0.1818182 0.1818182 0.1818182 0.1818182 0.6666667 0.6666667 1.0000000 0.1818182 Qualitative Predictions- “yes” “no” “no” “no” “yes” “no” “no” “yes” “no” “no” “no” “no” “yes” “yes” “yes” “yes” “no” “no” “no” “no” “yes” “yes” “yes” “no”

2.3. Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1 A. The naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1 is ‘0’,

---

2(4) Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

A. By the two predictions from bayes and naiveBayes theorems and taking the cutoff value 0.4 for the naiveBayes, it is clear that the two predictions are almost same with slight difference for the 24 rows that we have predicted.

3.Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%). 3(1) Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix. A. Reference Prediction no yes no 3203 5016 yes 2862 5793, Accuracy = 0.533

3(2) What is the overall error of the validation set? A. The overall error of validation set is 0.466

Problem Statement

The file accidentsFull.csv contains information on 42,183 actual automobile accidents in 2001 in the United States that involved one of three levels of injury: NO INJURY, INJURY, or FATALITY. For each accident, additional information is recorded, such as day of week, weather conditions, and road type. A firm might be interested in developing a system for quickly classifying the severity of an accident based on initial reports and associated data in the system (some of which rely on GPS-assisted reporting).

Our goal here is to predict whether an accident just reported will involve an injury (MAX_SEV_IR = 1 or 2) or will not (MAX_SEV_IR = 0). For this purpose, create a dummy variable called INJURY that takes the value “yes” if MAX_SEV_IR = 1 or 2, and otherwise “no.”

Data Import And Cleaning

Load the Required Libraries

```
library(caret)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: lattice
```

```
library(e1071)
```

data import which was in .csv format

```
accidents_data <- read.csv("C:\\Users\\rupes\\OneDrive\\Desktop\\Kent State University\\FML\\Ass
ignment 3\\accidentsFull.csv")

dim(accidents_data)
```

```
## [1] 42183    24
```

Create a Dummy variable “INJURY”

```
# create a new variable INJURY with "yes" or "no" by using the column MAX_SEV_IR
accidents_data$INJURY = ifelse(accidents_data$MAX_SEV_IR > 0, "yes", "no")
```

Questions

1. Using the information in this dataset, if an accident has just been reported and no further information is available, what should the prediction be? (INJURY = Yes or No?) Why?

```
# Accidents that are not Injured
accidents_I_N <- sum(accidents_data$INJURY == "no")


cat("No.of Accidents that are not Injured are", accidents_I_N,"\n")
```

```
## No.of Accidents that are not Injured are 20721
```

```
# Accidents that are Injured
accidents_I_Y <- sum(accidents_data$INJURY == "yes")

# Print the data
cat("No.of Accidents that are Injured are", accidents_I_Y,"\n")
```

```
## No.of Accidents that are Injured are 21462
```

From the above data, the accidents that are Injured was 21462 and the accidents that are not injured was 20721. by using the above information, if an accident has just been reported and no further information is available,I can predict that the reported accident is “INJURED” that means INJURY = YES.

Converting the variables to factors

```
# Converting variables of the dataset to factors
for (i in c(1:dim(accidents_data)[2])){
  accidents_data[,i] <- as.factor(accidents_data[,i])
}
```

---

2. Select the first 24 records in the dataset and look only at the response (INJURY) and the two predictors WEATHER_R and TRAF_CON_R. Create a pivot table that examines INJURY as a function of the two predictors for these 24 records. Use all three variables in the pivot table as rows/columns. 2(1). Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors. 2(2). Classify the 24 accidents using these probabilities and a cutoff of 0.5. 2(3). Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 2(4). Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

Create a dataframe by taking the first 24 columns and 3 columns “INJURY”, “WEATHER_R”, “TRAF_CON_R” from actual dataframe(accidents)

```
# Create a dataframe by taking the first 24 rows
accidentsdf24 <- accidents_data[1:24,c("INJURY", "WEATHER_R", "TRAF_CON_R")]
dim(accidentsdf24)
```

```
## [1] 24  3
```

create a pivot table from the above accidents24

```
# Create a pivot table using ftable function
data1 <- ftable(accidentsdf24) #ftable for creating pivot table
data2 <- ftable(accidentsdf24[,-1]) #pivot table by dropping the first column

# print the table
data1
```

```
##                 TRAF_CON_R 0 1 2
## INJURY WEATHER_R
## no     1                   3 1 1
##        2                   9 1 0
## yes    1                   6 0 0
##        2                   2 0 1
```

```
data2
```

```
##           TRAF_CON_R  0  1  2
## WEATHER_R
## 1                     9  1  1
## 2                    11  1  1
```

2.1 Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.

Considering Injury = yes and getting six possible combinations of the predictors.

```
#Probability when INJURY = YES
y1 <- data1[3,1] / data2[1,1] #WEATHER = 1, TRAFFIC = 0
y2 <- data1[4,1] / data2[2,1] #WEATHER = 2, TRAFFIC = 0
y3 <- data1[3,2] / data2[1,2] #WEATHER = 1, TRAFFIC = 1
y4 <- data1[4,2] / data2[2,2] #WEATHER = 2, TRAFFIC = 1
y5 <- data1[3,3] / data2[1,3] #WEATHER = 1, TRAFFIC = 2
y6 <- data1[4,3] / data2[2,3] #WEATHER = 2, TRAFFIC = 2

cat("when INJURY = YES the probalities are", "\n")
```

```
## when INJURY = YES the probalities are
```

```
c(y1, y2, y3, y4, y5, y6)
```

```
## [1] 0.6666667 0.1818182 0.0000000 0.0000000 0.0000000 1.0000000
```

```
#PROBABILITY when INJURY = NO
# Probability of INJURY = no, when
n1 <- data1[1,1] / data2[1,1] #WEATHER = 1, TRAFFIC = 0
n2 <- data1[2,1] / data2[2,1] #WEATHER = 2, TRAFFIC = 0
n3 <- data1[1,2] / data2[1,2] #WEATHER = 1, TRAFFIC = 1
n4 <- data1[2,2] / data2[2,2] #WEATHER = 2, TRAFFIC = 1
n5 <- data1[1,3] / data2[1,3] #WEATHER = 1, TRAFFIC = 2
n6 <- data1[2,3] / data2[2,3] #WEATHER = 2, TRAFFIC = 2


cat("when INJURY = NO the probabilities are", "\n")
```

```
## when INJURY = NO the probabilities are
```

```
c(n1, n2, n3, n4, n5, n6)
```

```
## [1] 0.3333333 0.8181818 1.0000000 1.0000000 1.0000000 0.0000000
```

2(2) Classify the 24 accidents using these probabilities and a cutoff of 0.5. Assigning the probabilities to the each of the 24rows.

```
# Considering the data from 0 to 24
probability_injury <- rep(0,24)

# use for loop considering iterations from 1 to 24
for(i in 1:24){
  if (accidentsdf24$WEATHER_R[i] == "1") {

      if (accidentsdf24$TRAF_CON_R[i]=="0"){
        probability_injury[i] = y1
      }

      else if (accidentsdf24$TRAF_CON_R[i]=="1") {
        probability_injury[i] = y3
      }

      else if (accidentsdf24$TRAF_CON_R[i]=="2") {
        probability_injury[i] = y5
      }
    }

    else {

      if (accidentsdf24$TRAF_CON_R[i]=="0"){
        probability_injury[i] = y2
      }

      else if (accidentsdf24$TRAF_CON_R[i]=="1") {
        probability_injury[i] = y4
      }

      else if (accidentsdf24$TRAF_CON_R[i]=="2") {
        probability_injury[i] = y6
      }
    }
  }

# Inserting the probabilities to the dataframe
accidentsdf24$probability_injury <- probability_injury

# Classifying the accidents by means of cutoff value 0.5
accidentsdf24$pred_prob <- ifelse(accidentsdf24$probability_injury > 0.5, "yes", "no")

accidentsdf24
```

```
##    INJURY WEATHER_R TRAF_CON_R probability_injury pred_prob
## 1     yes         1          0          0.6666667       yes
## 2      no         2          0          0.1818182        no
## 3      no         2          1          0.0000000        no
## 4      no         1          1          0.0000000        no
## 5      no         1          0          0.6666667       yes
## 6     yes         2          0          0.1818182        no
## 7      no         2          0          0.1818182        no
## 8     yes         1          0          0.6666667       yes
## 9      no         2          0          0.1818182        no
## 10     no         2          0          0.1818182        no
## 11     no         2          0          0.1818182        no
## 12     no         1          2          0.0000000        no
## 13    yes         1          0          0.6666667       yes
## 14     no         1          0          0.6666667       yes
## 15    yes         1          0          0.6666667       yes
## 16    yes         1          0          0.6666667       yes
## 17     no         2          0          0.1818182        no
## 18     no         2          0          0.1818182        no
## 19     no         2          0          0.1818182        no
## 20     no         2          0          0.1818182        no
## 21    yes         1          0          0.6666667       yes
## 22     no         1          0          0.6666667       yes
## 23    yes         2          2          1.0000000       yes
## 24    yes         2          0          0.1818182        no
```

Out of 24 rows there are 5 rows that the actual values of Injury from the dataset that does not matches with the predicted values. row5: the actual Injury was No, but the predicted was Yes row6: the actual Injury was Yes, but the predicted was No row14: the actual Injury was No, but the predicted was Yes row21: the actual Injury was No, but the predicted was Yes row14: the actual Injury was Yes, but the predicted was No

---

2.3 Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1. Computing manually the naive Bayes conditional probability

```
# Probability of getting Injured when WEATHER = 1
PIy_W1 <-  (data1[3,1] + data1[3,2] + data1[3,3]) / (data1[3,1] + data1[3,2] + data1[3,3] + data
1[4,1] + data1[4,2] + data1[4,3])
PIy_W1
```

```
## [1] 0.6666667
```

```
# Probability of getting Injured when TRAF_CON_R = 1
PIy_T1 <- (data1[3,2] + data1[4,2]) / (data1[3,1] + data1[3,2] + data1[3,3] + data1[4,1] + data1
[4,2] + data1[4,3])

# Probability of getting Injured
PIy <- (data1[3,1] + data1[3,2] + data1[3,3] + data1[4,1] + data1[4,2] + data1[4,3])/24

# Probability of not getting Injured when WEATHER_R = 1
PIn_W1 <- (data1[1,1] + data1[1,2] + data1[1,3]) / (data1[1,1] + data1[1,2] + data1[1,3] + data1
[2,1] + data1[2,2] + data1[2,3])

# Probability of not getting Injured when TRAF_CON_R = 1
PIn_T1 <- (data1[1,2] + data1[2,2]) / (data1[1,1] + data1[1,2] + data1[1,3] + data1[2,1] + data1
[2,2] + data1[2,3])

# Probability of not getting Injured
PIn <- (data1[1,1] + data1[1,2] + data1[1,3] + data1[2,1] + data1[2,2] + data1[2,3])/24

# Probability of getting Injured when WEATHER_R = 1 and TRAF_CON_R = 1
PW1_T1 <- (PIy_W1 * PIy_T1 * PIy)/ ((PIy_W1 * PIy_T1 * PIy) + (PIn_W1 * PIn_T1 * PIn))

cat("The naive Bayes conditional probability of an injury WHEN WEATHER = 1 and TRAFFIC = 1 is",
PW1_T1)
```

```
## The naive Bayes conditional probability of an injury WHEN WEATHER = 1 and TRAFFIC = 1 is 0
```

2(4). Run a naive Bayes classifier on the 24 records and two predictors. Check the model output to obtain probabilities and classifications for all 24 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (= ordering) of observations equivalent?

Training and Predicting the data

```
# training the naiveBayes model by considering the predictors, Traffic and weather
nb1 <- naiveBayes(INJURY ~ TRAF_CON_R + WEATHER_R, data = accidentsdf24)

# Predicting the data using naiveBayes model
nbT <- predict(nb1, newdata = accidentsdf24, type = "raw")

# Inserting the newly predicted data to  accidents24 dataframe
accidentsdf24$nbpred_probability <- nbT[,2] # Transfer the "Yes" nb prediction


# Consider cutoff value 0.4 for naiveBayes predictions
accidentsdf24$nbpred_probability_condition <- ifelse(accidentsdf24$nbpred_probability>0.4, "ye
s", "no") #if probability was greater than 0.4 the Injury will be yes
accidentsdf24
```

```
##    INJURY WEATHER_R TRAF_CON_R probability_injury pred_prob nbpred_probability
## 1     yes         1          0          0.6666667       yes        0.571428571
## 2      no         2          0          0.1818182        no        0.250000000
## 3      no         2          1          0.0000000        no        0.002244949
## 4      no         1          1          0.0000000        no        0.008919722
## 5      no         1          0          0.6666667       yes        0.571428571
## 6     yes         2          0          0.1818182        no        0.250000000
## 7      no         2          0          0.1818182        no        0.250000000
## 8     yes         1          0          0.6666667       yes        0.571428571
## 9      no         2          0          0.1818182        no        0.250000000
## 10     no         2          0          0.1818182        no        0.250000000
## 11     no         2          0          0.1818182        no        0.250000000
## 12     no         1          2          0.0000000        no        0.666666667
## 13    yes         1          0          0.6666667       yes        0.571428571
## 14     no         1          0          0.6666667       yes        0.571428571
## 15    yes         1          0          0.6666667       yes        0.571428571
## 16    yes         1          0          0.6666667       yes        0.571428571
## 17     no         2          0          0.1818182        no        0.250000000
## 18     no         2          0          0.1818182        no        0.250000000
## 19     no         2          0          0.1818182        no        0.250000000
## 20     no         2          0          0.1818182        no        0.250000000
## 21    yes         1          0          0.6666667       yes        0.571428571
## 22     no         1          0          0.6666667       yes        0.571428571
## 23    yes         2          2          1.0000000       yes        0.333333333
## 24    yes         2          0          0.1818182        no        0.250000000
##    nbpred_probability_condition
## 1                           yes
## 2                            no
## 3                            no
## 4                            no
## 5                           yes
## 6                            no
## 7                            no
## 8                           yes
## 9                            no
## 10                           no
## 11                           no
## 12                          yes
## 13                          yes
## 14                          yes
## 15                          yes
## 16                          yes
## 17                           no
## 18                           no
## 19                           no
## 20                           no
## 21                          yes
## 22                          yes
## 23                           no
## 24                           no
```

```
#Loading the klaR package for Naive Bayes
library(klaR)
```

```
## Loading required package: MASS
```

```
Dset <- INJURY ~ TRAF_CON_R + WEATHER_R
accidentsdf24$INJURY <- as.factor(accidentsdf24$INJURY)

nb2 <- NaiveBayes(Dset,data = accidentsdf24)

#predicting data
predict(nb2, newdata = accidentsdf24[,c("INJURY", "WEATHER_R", "TRAF_CON_R")])
```

```
## $class
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20
## yes  no  no  no yes  no  no yes  no  no  no yes yes yes yes yes  no  no  no  no
##  21  22  23  24
## yes yes  no  no
## Levels: no yes
##
## $posterior
##           no         yes
## 1  0.4285714 0.571428571
## 2  0.7500000 0.250000000
## 3  0.9977551 0.002244949
## 4  0.9910803 0.008919722
## 5  0.4285714 0.571428571
## 6  0.7500000 0.250000000
## 7  0.7500000 0.250000000
## 8  0.4285714 0.571428571
## 9  0.7500000 0.250000000
## 10 0.7500000 0.250000000
## 11 0.7500000 0.250000000
## 12 0.3333333 0.666666667
## 13 0.4285714 0.571428571
## 14 0.4285714 0.571428571
## 15 0.4285714 0.571428571
## 16 0.4285714 0.571428571
## 17 0.7500000 0.250000000
## 18 0.7500000 0.250000000
## 19 0.7500000 0.250000000
## 20 0.7500000 0.250000000
## 21 0.4285714 0.571428571
## 22 0.4285714 0.571428571
## 23 0.6666667 0.333333333
## 24 0.7500000 0.250000000
```

3. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%). 3(1) Run a naive Bayes classifier on the complete training set with the relevant predictors (and INJURY as the response). Note that all predictors are categorical. Show the confusion matrix.

Splitting the Data into 60% training and 40% validation.

```
set.seed(1)

train_set <- sample(row.names(accidents_data), 0.6*dim(accidents_data)[1]) # 60% training data
valid_set <- setdiff(row.names(accidents_data), train_set) # 40% validation data

train_data <- accidents_data[train_set,]
valid_data <- accidents_data[valid_set,]

#Defining what variables to be used

variables <- c("INJURY", "HOUR_I_R", "ALIGN_I" ,"WRK_ZONE",  "WKDY_I_R",
          "INT_HWY",  "LGTCON_I_R", "PROFIL_I_R", "SPD_LIM", "SUR_COND",
          "TRAF_CON_R",   "TRAF_WAY",   "WEATHER_R")

naive_prediction <- naiveBayes(INJURY~.,data = train_data[,variables])

naive_prediction
```

```
##
## Naive Bayes Classifier for Discrete Predictors
##
## Call:
## naiveBayes.default(x = X, y = Y, laplace = laplace)
##
## A-priori probabilities:
## Y
##        no       yes
## 0.4939745 0.5060255
##
## Conditional probabilities:
##      HOUR_I_R
## Y             0         1
##   no  0.5689490 0.4310510
##   yes 0.5703131 0.4296869
##
##      ALIGN_I
## Y             1         2
##   no  0.8712206 0.1287794
##   yes 0.8652300 0.1347700
##
##      WRK_ZONE
## Y              0          1
##   no  0.97664374 0.02335626
##   yes 0.97727805 0.02272195
##
##      WKDY_I_R
## Y             0         1
##   no  0.2194049 0.7805951
##   yes 0.2381510 0.7618490
##
##      INT_HWY
## Y                0            1            9
##   no  0.8513837786 0.1481362982 0.0004799232
##   yes 0.8593737800 0.1397673147 0.0008589053
##
##      LGTCON_I_R
## Y             1         2         3
##   no  0.6870101 0.1251000 0.1878899
##   yes 0.7014914 0.1096275 0.1888811
##
##      PROFIL_I_R
## Y             0         1
##   no  0.7531595 0.2468405
##   yes 0.7633326 0.2366674
##
##      SPD_LIM
## Y                5           10           15           20           25
##   no  0.0000799872 0.0004799232 0.0043992961 0.0085586306 0.1121420573
##   yes 0.0001561646 0.0003123292 0.0040602795 0.0039041149 0.0906535488
##      SPD_LIM
```

```
## Y               30           35           40           45           50
##   no  0.0860662294 0.1896496561 0.0962246041 0.1553351464 0.0407934730
##   yes 0.0860466932 0.2123057703 0.1068946670 0.1574139143 0.0394315609
##      SPD_LIM
## Y               55           60           65           70           75
##   no  0.1590145577 0.0355143177 0.0645496721 0.0409534474 0.0062390018
##   yes 0.1549152807 0.0430233466 0.0621535098 0.0311548372 0.0075739830
##
##      SUR_COND
## Y               1           2           3           4           9
##   no  0.774196129 0.176931691 0.016717325 0.028155495 0.003999360
##   yes 0.815725775 0.151245413 0.010697275 0.016709612 0.005621926
##
##      TRAF_CON_R
## Y             0         1         2
##   no  0.6566149 0.1902096 0.1531755
##   yes 0.6213009 0.2191770 0.1595221
##
##      TRAF_WAY
## Y              1          2          3
##   no  0.57998720 0.36690130 0.05311150
##   yes 0.56063090 0.39743890 0.04193019
##
##      WEATHER_R
## Y             1         2
##   no  0.8390657 0.1609343
##   yes 0.8744437 0.1255563
```

3(2) What is the overall error of the validation set?

```
confusn_Matrix =  confusionMatrix(valid_data$INJURY, predict(naive_prediction, valid_data[, vari
ables]), positive = "yes")

confusn_Matrix
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   no  yes
##        no  3203 5016
##        yes 2862 5793
##
##               Accuracy : 0.5331
##                 95% CI : (0.5256, 0.5407)
##     No Information Rate : 0.6406
##     P-Value [Acc > NIR] : 1
##
##                  Kappa : 0.0594
##
##  Mcnemar's Test P-Value : <2e-16
##
##             Sensitivity : 0.5359
##             Specificity : 0.5281
##          Pos Pred Value : 0.6693
##          Neg Pred Value : 0.3897
##              Prevalence : 0.6406
##          Detection Rate : 0.3433
##    Detection Prevalence : 0.5129
##       Balanced Accuracy : 0.5320
##
##        'Positive' Class : yes
##
```

#Overall error of the validation set

```
overall_error_rate = 1 - confusn_Matrix$overall["Accuracy"]
cat("The Overall Error is : ", overall_error_rate)
```

```
## The Overall Error is :  0.4668721
```