FML_Assignment3_NIHA

Niharika Matsa

2023-10-15

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."

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?
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 12 records. Use all three variables in the pivot table as rows/columns.

- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.
- Classify the 24 accidents using these probabilities and a cutoff of 0.5.
- Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1.
- 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?

3. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%).

- 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.
- What is the overall error of the validation set?

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?

If there is no additional information giventhen we can rely on past data, that is probability can be used to predict if there is any injury or not.

Data Input and Cleaning

Load the required libraries and read the input file

```
library(e1071)
library(caret)
```

```
## Loading required package: ggplot2

## Loading required package: lattice
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(klaR)
```

```
## Loading required package: MASS

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
accidents = read.csv("C:/Users/Krupa shetty/Downloads/accidentsFull.csv")

accidents$INJURY = ifelse(accidents$MAX_SEV_IR>0,"yes","no")

head(accidents)
```

```
##   HOUR_I_R ALCHL_I ALIGN_I STRATUM_R WRK_ZONE WKDY_I_R INT_HWY LGTCON_I_R
## 1        0       2       2         1        0        1       0          3
## 2        1       2       1         0        0        1       1          3
## 3        1       2       1         0        0        1       0          3
## 4        1       2       1         1        0        0       0          3
## 5        1       1       1         0        0        1       0          3
## 6        1       2       1         1        0        1       0          3
##   MANCOL_I_R PED_ACC_R RELJCT_I_R REL_RWY_R PROFIL_I_R SPD_LIM SUR_COND
## 1          0         0          1         0          1      40        4
## 2          2         0          1         1          1      70        4
## 3          2         0          1         1          1      35        4
## 4          2         0          1         1          1      35        4
## 5          2         0          0         1          1      25        4
## 6          0         0          1         0          1      70        4
##   TRAF_CON_R TRAF_WAY VEH_INVL WEATHER_R INJURY_CRASH NO_INJ_I PRPTYDMG_CRASH
## 1          0        3        1         1            1        1              0
## 2          0        3        2         2            0        0              1
## 3          1        2        2         2            0        0              1
## 4          1        2        2         1            0        0              1
## 5          0        2        3         1            0        0              1
## 6          0        2        1         2            1        1              0
##   FATALITIES MAX_SEV_IR INJURY
## 1          0          1    yes
## 2          0          0     no
## 3          0          0     no
## 4          0          0     no
## 5          0          0     no
## 6          0          1    yes
```

```
# Convert variables to factor
for (i in c(1:dim(accidents)[2])){
  accidents[,i] = as.factor(accidents[,i])
}
head(accidents,n=24)
```

```
##    HOUR_I_R ALCHL_I ALIGN_I STRATUM_R WRK_ZONE WKDY_I_R INT_HWY LGTCON_I_R
## 1         0       2       2         1        0        1       0          3
## 2         1       2       1         0        0        1       1          3
## 3         1       2       1         0        0        1       0          3
## 4         1       2       1         1        0        0       0          3
## 5         1       1       1         0        0        1       0          3
## 6         1       2       1         1        0        1       0          3
## 7         1       2       1         0        0        1       1          3
## 8         1       2       1         1        0        1       0          3
## 9         1       2       1         1        0        1       0          3
## 10        0       2       1         0        0        0       0          3
## 11        1       2       1         0        0        1       0          3
## 12        1       2       1         1        0        1       0          3
## 13        1       2       1         1        0        1       0          3
## 14        1       2       2         0        0        1       0          3
## 15        1       2       2         1        0        1       0          3
## 16        1       2       2         1        0        1       0          3
## 17        1       2       1         1        0        1       0          3
## 18        1       2       1         1        0        0       0          3
## 19        1       2       1         1        0        1       0          3
## 20        1       2       1         0        0        1       0          3
## 21        1       2       1         1        0        1       0          3
## 22        1       2       2         0        0        1       0          3
## 23        1       2       1         0        0        1       0          3
## 24        1       2       1         1        0        1       9          3
##    MANCOL_I_R PED_ACC_R RELJCT_I_R REL_RWY_R PROFIL_I_R SPD_LIM SUR_COND
## 1           0         0          1         0          1      40        4
## 2           2         0          1         1          1      70        4
## 3           2         0          1         1          1      35        4
## 4           2         0          1         1          1      35        4
## 5           2         0          0         1          1      25        4
## 6           0         0          1         0          1      70        4
## 7           0         0          0         0          1      70        4
## 8           0         0          0         0          1      35        4
## 9           0         0          1         0          1      30        4
## 10          0         0          1         0          1      25        4
## 11          0         0          0         0          1      55        4
## 12          2         0          0         1          1      40        4
## 13          1         0          0         1          1      40        4
## 14          0         0          0         0          1      25        4
## 15          0         0          0         0          1      35        4
## 16          0         0          0         0          1      45        4
## 17          0         0          0         0          1      20        4
## 18          0         0          0         0          1      50        4
## 19          0         0          0         0          1      55        4
## 20          0         0          1         1          1      55        4
## 21          0         0          1         0          0      45        4
## 22          0         0          1         0          0      65        4
## 23          0         0          0         0          0      65        4
## 24          2         0          1         1          0      55        4
##    TRAF_CON_R TRAF_WAY VEH_INVL WEATHER_R INJURY_CRASH NO_INJ_I PRPTYDMG_CRASH
## 1           0        3        1         1            1        1              0
## 2           0        3        2         2            0        0              1
## 3           1        2        2         2            0        0              1
## 4           1        2        2         1            0        0              1
## 5           0        2        3         1            0        0              1
## 6           0        2        1         2            1        1              0
## 7           0        2        1         2            0        0              1
## 8           0        1        1         1            1        1              0
## 9           0        1        1         2            0        0              1
## 10          0        1        1         2            0        0              1
## 11          0        1        1         2            0        0              1
## 12          2        1        2         1            0        0              1
## 13          0        1        4         1            1        2              0
## 14          0        1        1         1            0        0              1
## 15          0        1        1         1            1        1              0
## 16          0        1        1         1            1        1              0
## 17          0        1        1         2            0        0              1
## 18          0        1        1         2            0        0              1
## 19          0        1        1         2            0        0              1
## 20          0        1        1         2            0        0              1
## 21          0        3        1         1            1        1              0
## 22          0        3        1         1            0        0              1
## 23          2        2        1         2            1        2              0
## 24          0        2        2         2            1        1              0
##    FATALITIES MAX_SEV_IR INJURY
## 1           0          1    yes
## 2           0          0     no
## 3           0          0     no
## 4           0          0     no
## 5           0          0     no
## 6           0          1    yes
## 7             0          0     no
## 8             0          1    yes
## 9             0          0     no
## 10            0          0     no
## 11            0          0     no
## 12            0          0     no
## 13            0          1    yes
## 14            0          0     no
## 15            0          1    yes
## 16            0          1    yes
## 17            0          0     no
## 18            0          0     no
## 19            0          0     no
## 20            0          0     no
## 21            0          1    yes
## 22            0          0     no
## 23            0          1    yes
## 24            0          1    yes
```

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?

If there is no additional information giventhen we can rely on past data, that is probability can be used to predict if there is any injury or not.

Probability of injury=yes

```
prob_yes<- accidents%>% filter(accidents$INJURY=="yes")%>%summarise(count= n())
inj_yes<-  prob_yes/nrow(accidents)
inj_yes
```

```
##       count
## 1 0.5087832
```

Probability of injury=No

```
prob_no<- accidents%>% filter(accidents$INJURY=="no")%>%summarise(count= n())
inj_no<-  prob_no/nrow(accidents)
inj_no
```

```
##       count
## 1 0.4912168
```

As the probability of injury=no is less than probability of injury=yes, so there is more chances that people get injured when accidents occur.

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 12 records. Use all three variables in the pivot table as rows/columns.

- Compute the exact Bayes conditional probabilities of an injury (INJURY = Yes) given the six possible combinations of the predictors.
- Classify the 24 accidents using these probabilities and a cutoff of 0.5.
- Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and TRAF_CON_R = 1.
- 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?

```
accidents24 = accidents[1:24,c("INJURY","WEATHER_R","TRAF_CON_R")]
#head(accidents24)
```

```
dt1 = ftable(accidents24)
dt2 = ftable(accidents24[,-1]) # print table only for conditions
head(dt1)
```

```
##
##                         "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
```

```
head(dt2)
```

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

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 12 records. Use all three variables in the pivot table as rows/columns.

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

```
# Injury = yes
p1 = dt1[3,1] / dt2[1,1] # Injury, Weather=1 and Traf=0
p2 = dt1[4,1] / dt2[2,1] # Injury, Weather=2, Traf=0
p3 = dt1[3,2] / dt2[1,2] # Injury, W=1, T=1
p4 = dt1[4,2] / dt2[2,2] # I, W=2,T=1
p5 = dt1[3,3] / dt2[1,3] # I, W=1,T=2
p6 = dt1[4,3]/ dt2[2,3] #I,W=2,T=2

# Injury = no
n1 = dt1[1,1] / dt2[1,1] # Weather=1 and Traf=0
n2 = dt1[2,1] / dt2[2,1] # Weather=2, Traf=0
n3 = dt1[1,2] / dt2[1,2] # W=1, T=1
```

```
n4 = dt1[2,2] / dt2[2,2] # W=2,T=1
n5 = dt1[1,3] / dt2[1,3] # W=1,T=2
n6 = dt1[2,3] / dt2[2,3] # W=2,T=2
print(c(p1,p2,p3,p4,p5,p6))
```

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

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

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

1.Probability of Injury is yes for Weather = 1 and Traffic = 0 is 0.6666667 2.Probability of Injury is yes for Weather = 1 and Traffic = 1 is 0.0000000 3.Probability of Injury is yes for Weather = 1 and Traffic = 2 is 0.0000000 4.Probability of Injury is no for Weather = 1 and Traffic = 0 is 0.3333333 5.Probability of Injury is no for Weather = 1 and Traffic = 1 is 1.0000000 6.Probability of Injury is no for Weather = 1 and Traffic = 2 is 1.0000000 7.Probability of Injury is yes for Weather = 2 and Traffic = 0 is 0.1818182 8.Probability of Injury is yes for Weather = 2 and Traffic = 1 is 0.0000000 9.Probability of Injury is yes for Weather = 2 and Traffic = 2 is 1.0000000 10.Probability of Injury is no for Weather = 2 and Traffic = 0 is 0.8181818 11.Probability of Injury is no for Weather = 2 and Traffic = 1 is 1.0000000 12.Probability of Injury is no for Weather = 2 and Traffic = 2 is 0.0000000

2. Let us now compute

- Classify the 24 accidents using these probabilities and a cutoff of 0.5.

```
prob.inj = rep(0,24)

for (i in 1:24) {
  print(c(accidents24$WEATHER_R[i],accidents24$TRAF_CON_R[i]))
    if (accidents24$WEATHER_R[i] == "1") {
      if (accidents24$TRAF_CON_R[i]=="0"){
        prob.inj[i] = p1
      }
      else if (accidents24$TRAF_CON_R[i]=="1") {
        prob.inj[i] = p3
      }
      else if (accidents24$TRAF_CON_R[i]=="2") {
        prob.inj[i] = p5
      }
    }
    else {
      if (accidents24$TRAF_CON_R[i]=="0"){
        prob.inj[i] = p2
      }
      else if (accidents24$TRAF_CON_R[i]=="1") {
        prob.inj[i] = p4
      }
      else if (accidents24$TRAF_CON_R[i]=="2") {
        prob.inj[i] = p6
      }
    }
  }
```

```
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 1
## Levels: 1 2 0
## [1] 1 1
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 1 2
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 1 0
## Levels: 1 2 0
## [1] 2 2
## Levels: 1 2 0
## [1] 2 0
## Levels: 1 2 0
```

```
accidents24$prob.inj = prob.inj

accidents24$pred.prob = ifelse(accidents24$prob.inj>0.5, "yes", "no")
```

Compute manually the naive Bayes conditional probability of an injury given WEATHER_R = 1 and

TRAF_CON_R = 1.

```r
df = accidents24[accidents24$WEATHER_R == "1" & accidents24$TRAF_CON_R == "1", ]

probability = sum(df$INJURY == "yes") / nrow(df)

probability
```

```
## [1] 0
```

```r
#The probability is 0 for injury is yes and weather is 1 and traffic is 1.
```

```r
dataset = data.frame(WEATHER_R = "1", TRAF_CON_R = "1")
bayes_naive= naiveBayes(INJURY ~ TRAF_CON_R + WEATHER_R,
                 data = accidents24)

predict = predict(bayes_naive, newdata = dataset, type = "raw")

predict
```

```
##              no         yes
## [1,] 0.9910803 0.008919722
```

The probability of injury is yes is 0.008919722 for weather=1 and traffic= 1.

2.

- 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?

```r
bayes_naive= naiveBayes(INJURY ~ TRAF_CON_R + WEATHER_R,
                 data = accidents24)

naive = predict(bayes_naive, newdata = accidents24)
# accidents24$nbpred.prob = naive[,2] # Transfer the "Yes" nb prediction

naive
```

```
##  [1] yes no  no  no  yes no  no  yes no  no  no  yes yes yes yes yes no  no  no
## [20] no  yes yes no  no
## Levels: no yes
```

```r
cutoff = 0.5

bayes = ifelse(c(p1, p2, p3, p4, p5, p6) > cutoff, "yes", "no")

final = data.frame(
  "Bayes" = bayes,
  "Probability" = naive)
```

```
naive_classifications = bayes == naive

order = order(-as.numeric(c(p1, p2, p3, p4, p5, p6))) == order(-as.numeric(naive))

head(final)
```

```
##   Bayes Probability
## 1   yes         yes
## 2    no          no
## 3    no          no
## 4    no          no
## 5    no         yes
## 6   yes          no
```

```
cat("Are the resulting classifications equivalent? ", all(naive_classifications), "\n")
```

```
## Are the resulting classifications equivalent?  FALSE
```

```
cat("Is the ranking of observations equivalent? ", all(order), "\n")
```

```
## Is the ranking of observations equivalent?  FALSE
```

For all records, the exact Bayes and Naive Bayes classifications do not agree, and neither do the rankings of the observations for the two approaches. This shows that in this particular dataset and feature set, the Naive Bayes classifier might produce different predictions from the exact Bayes classifier.

3. Let us now return to the entire dataset. Partition the data into training (60%) and validation (40%).

- 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.
- What is the overall error of the validation set?

```
###setting the seed
set.seed(1)
###reading the data
accidents_new = read.csv("C:/Users/Krupa shetty/Downloads/accidentsFull.csv")


accidents_new$INJURY = ifelse(accidents_new$MAX_SEV_IR>0,1,0)

for (i in c(1:dim(accidents_new)[2])){
  accidents[,i] = as.factor(accidents_new[,i])
}

###splitting the data
train = sample(row.names(accidents_new), 0.6*dim(accidents_new)[1])

valid = setdiff(row.names(accidents_new), train)
```

```r
train_data = accidents_new[train,]

valid_data = accidents_new[valid,]

###using naive bayes on data
model = naiveBayes(INJURY ~ ., data = train_data)

prediction = predict(model, valid_data)


###Confusion matrix
matrix = confusionMatrix(prediction, as.factor(valid_data$INJURY))

print(matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 8219  205
##          1    0 8450
##
##                Accuracy : 0.9879
##                  95% CI : (0.9861, 0.9894)
##     No Information Rate : 0.5129
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9757
##
##  Mcnemar’s Test P-Value : < 2.2e-16
##
##             Sensitivity : 1.0000
##             Specificity : 0.9763
##          Pos Pred Value : 0.9757
##          Neg Pred Value : 1.0000
##              Prevalence : 0.4871
##          Detection Rate : 0.4871
##    Detection Prevalence : 0.4992
##       Balanced Accuracy : 0.9882
##
##        ’Positive’ Class : 0
##
```

```r
###Error
error_rate = 1 - matrix$overall["Accuracy"]

error_rate
```

```
##   Accuracy
## 0.01214887
```

The model does well when it comes to prediction because the overall error is 0.01214887. The model has great sensitivity, specificity,and accuracy. as the error is low, the model predictions are almost accurate.