FML Assignment 3 - Predicting Accident Injuries

HARI VINAYAK

2023-10-16

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.”

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? 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. 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? 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?

#Summary of Results:

We will predict whether an accident involves an injury (INJURY = Yes) or not (INJURY = No) based on the given dataset. First, we’ll examine the data for the first 24 records, then compute exact Bayes conditional probabilities, and later compare the results with a naive Bayes classifier. Finally, we’ll partition the data, run a naive Bayes classifier on the training set, and calculate the overall error on the validation set.

There is a 47.7% of a injury happening.

Solution

Step 1: Data Exploration and Transformation

```
library(e1071)

data <- read.csv("AccidentsFull.csv")

data$INJURY <- ifelse(data$MAX_SEV_IR %in% c(1, 2), "Yes", "No")

subset_data <- data[1:24, c("INJURY", "WEATHER_R", "TRAF_CON_R")]

pivot_table <- table(subset_data$INJURY, subset_data$WEATHER_R, subset_data$TRAF_CON_
R)

pivot_table
```

```
## , ,  = 0
##
##
##       1 2
##   No  3 9
##   Yes 6 2
##
## , ,  = 1
##
##
##       1 2
##   No  1 1
##   Yes 0 0
##
## , ,  = 2
##
##
##       1 2
##   No  1 0
##   Yes 0 1
```

Step 2: Exact Bayes Conditional Probabilities

For the exact Bayes conditional probabilities, we need to calculate the probabilities for each combination of predictors. We'll use these probabilities to classify the 24 accidents.

```
total_cases <- nrow(subset_data)

# Probability of INJURY = Yes
prob_injury_yes <- sum(subset_data$INJURY == "Yes") / total_cases

# Probability of INJURY = No
prob_injury_no <- sum(subset_data$INJURY == "No") / total_cases

# Calculate conditional probabilities for all combinations
# For example, P(INJURY = Yes | WEATHER_R = 1, TRAF_CON_R = 1)
prob_injury_yes_weather1_traf1 <- sum(subset_data$INJURY == "Yes" & subset_data$WEATH
ER_R == 1 & subset_data$TRAF_CON_R == 1) / total_cases
# Similarly, calculate for other combinations

# Classification using 0.5 cutoff
classified <- ifelse(prob_injury_yes >= 0.5, "Yes", "No")
classified
```

```
## [1] "No"
```

Step 3: Manual Naive Bayes Conditional Probability

Let's calculate the manual Naive Bayes conditional probability for WEATHER_R = 1 and TRAF_CON_R = 1.

```
prob_weather1_given_injury_yes <- sum(subset_data$WEATHER_R == 1 & subset_data$INJURY
== "Yes") / sum(subset_data$INJURY == "Yes")

# Probability of TRAF_CON_R = 1 given INJURY = Yes
prob_traf1_given_injury_yes <- sum(subset_data$TRAF_CON_R == 1 & subset_data$INJURY =
= "Yes") / sum(subset_data$INJURY == "Yes")

# Naive Bayes conditional probability for INJURY = Yes
naive_bayes_prob_injury_yes <- prob_weather1_given_injury_yes * prob_traf1_given_inju
ry_yes

naive_bayes_prob_injury_yes
```

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

Step 4: Naive Bayes Classifier for 24 Records

Now, let's run a Naive Bayes classifier on the 24 records.

```
# Load necessary libraries
library(e1071)

# Create a dataframe for classification
classification_data <- subset_data[c("WEATHER_R", "TRAF_CON_R")]
classification_data$INJURY <- as.factor(subset_data$INJURY)

# Run Naive Bayes classifier
naive_bayes_model <- naiveBayes(INJURY ~ ., data = classification_data)

# Predict probabilities and classifications
predictions <- predict(naive_bayes_model, classification_data, type = "raw")
classified_naive_bayes <- ifelse(predictions[, "Yes"] >= 0.5, "Yes", "No")

# Compare with exact Bayes classification
identical(classified, classified_naive_bayes)
```

```
## [1] FALSE
```

Step 5: Partition Data and Calculate Error on Validation Set

```
# Set a random seed for reproducibility
set.seed(123)

# Create a random index for data partition
index <- sample(1:nrow(data), nrow(data))

# 60% of data for training, 40% for validation
train_data <- data[index[1:round(0.6 * nrow(data))], ]
validation_data <- data[index[round(0.6 * nrow(data)) + 1:nrow(data)], ]

# Run Naive Bayes on the training set
naive_bayes_model_full <- naiveBayes(INJURY ~ WEATHER_R + TRAF_CON_R, data = train_da
ta)

# Make predictions on the validation set
validation_predictions <- predict(naive_bayes_model_full, newdata = validation_data,
type = "raw")

# Calculate the confusion matrix
confusion_matrix <- table(Actual = validation_data$INJURY, Predicted = ifelse(validat
ion_predictions[, "Yes"] >= 0.5, "Yes", "No"))

# Calculate overall error
overall_error <- 1 - sum(diag(confusion_matrix)) / sum(confusion_matrix)
overall_error
```

```
## [1] 0.4773899
```

```
#show the first few values
head(overall_error)
```

```
## [1] 0.4773899
```