HW4

February 25, 2024

Problem 4.1 (Automobile Accidents)

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."

a. 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?

b. Select the first 12 records in the dataset and look only at the response (`INJURY`) and the two predictors `WEATHER_R` and `TRAF_CON_R`.

i. 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.

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

iii. Classify the 12 accidents using these probabilities and a cutoff of 0.5.

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

v. Run a naive Bayes classifier on the 12 records and 2 predictors using scikitlearn. Check the model output to obtain probabilities and classifications for all 12 records. Compare this to the exact Bayes classification. Are the resulting classifications equivalent? Is the ranking (=ordering) of observations equivalent?

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

i. Assuming that no information or initial reports about the accident itself are available at the time of prediction (only location characteristics, weather conditions, etc.), which predictors can we include in the analysis? (See the data descriptions below)

ii. 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.

iii. What is the overall error for the validation set?

iv. What is the percent improvement relative to the naive rule (using the validation set)?

v. Examine the conditional probabilities in the pivot tables. Why do we get a probability of zero for `P(INJURY = No j SPD_LIM = 5)`?

Data Description

These data, from the U.S. Bureau of Transportation Statistics, can be used to predict whether an accident will result in injuries or fatalities, based on predictors such as alcohol involvement, time of day, road condition, etc. Such a prediction system could be used to prioritize responder resources at the time of the report.

Source: US Dept. of Transportation, Bureau of Transportation Statistics, "TranStats," (www.transtats.bts.gov – select "databases" then "General Estimate System (GES))

Note: TranStats reports both variables with missing data, and their derived counterparts with imputed values filled in, denoted by an "I" at the end. Only one variant (the original or the derived) is included here. An "R" at the end of the variable name indicates that the Transtats variable has been collapsed into fewer categories for analysis purposes. Data are for the year 2001.

[1]:
```
from IPython.display import Image
Image(filename='images/acc1.png', width=600)
```

[1]:

	Variables	
1	HOUR_I_R	1=rush hour, 0=not (rush = 6-9 am, 4-7 pm)
2	ALCOHOL_I	Alcohol involved = 1, not involved = 2
3	ALIGN_I	1 = straight, 2 = curve
4	STRATUM_R	1= NASS Crashes Involving At Least One Passenger Vehicle, i.e., A Passenger Car, Sport Utility Vehicle, Pickup Truck Or Van) Towed Due To Damage From The Crash Scene And No Medium Or Heavy Trucks Are Involved. 0=not
5	WRK_ZONE	1= yes, 0= no
6	WKDY_I_R	1=weekday, 0=weekend
7	INT_HWY	Interstate? 1=yes, 0= no
8	LGTCON_I_R	Light conditions - 1=day, 2=dark (including dawn/dusk), 3=dark, but lighted,4=dawn or dusk
9	MAN_COL_I	0=no collision, 1=head-on, 2=other form of collision
10	PED_ACC_R	1=pedestrian/cyclist involved, 0=not

[2]:
```
Image(filename='images/acc2.png', width=600)
```

[2]:

11	REL_JCT_I_R	1=accident at intersection/interchange, 0=not at intersection
12	REL_RWY_R	1=accident on roadway, 0=not on roadway
13	PROFIL_I_R	1= level, 0=other
14	SPD_LIM	Speed limit, miles per hour
15	SUR_CON	Surface conditions (1=dry, 2=wet, 3=snow/slush, 4=ice, 5=sand/dirt/oil, 8=other, 9=unknown)
16	TRAF_CON_R	Traffic control device: 0=none, 1=signal, 2=other (sign, officer …)
17	TRAF_WAY	1=two-way traffic, 2=divided hwy, 3=one-way road
18	VEH_INVL	Number of vehicles involved
19	WEATHER_R	1=no adverse conditions, 2=rain, snow or other adverse condition
20	NO_INJ_I	Number of injuries
21	PRPTYDMG_CRASH	1=property damage, 2=no property damage
22	FATALITIES	1= yes, 0= no
23	MAX_SEV_IR	0=no injury, 1=non-fatal inj., 2=fatal inj.

[3]:
```
import pandas as pd

df = pd.read_csv('../data/accidentsFull.csv')
df.head()
```

[3]:

	HOUR_I_R	ALCHL_I	ALIGN_I	STRATUM_R	WRK_ZONE	WKDY_I_R	INT_HWY	\
0	0	2	2	1	0	1	0	
1	1	2	1	0	0	1	1	
2	1	2	1	0	0	1	0	
3	1	2	1	1	0	0	0	
4	1	1	1	0	0	1	0	

	LGTCON_I_R	MANCOL_I_R	PED_ACC_R	…	SUR_COND	TRAF_CON_R	TRAF_WAY	\
0	3	0	0	…	4	0	3	
1	3	2	0	…	4	0	3	
2	3	2	0	…	4	1	2	
3	3	2	0	…	4	1	2	
4	3	2	0	…	4	0	2	

	VEH_INVL	WEATHER_R	INJURY_CRASH	NO_INJ_I	PRPTYDMG_CRASH	FATALITIES	\
0	1	1	1	1	0	0	
1	2	2	0	0	1	0	
2	2	2	0	0	1	0	
3	2	1	0	0	1	0	
4	3	1	0	0	1	0	

	MAX_SEV_IR
0	1
1	0
2	0

```
3           0
4           0

[5 rows x 24 columns]
```