Car Crash Data Analysis

Borja Calvo

Abstract

During the course we have presented different tools build probabilistic models from data, and we have seen how these tools can be used in R. In this project these tools have to be used to analyse a real dataset of car crashes. This dataset contains a number of variables related with the conditions under which car crashes happend in a certain road in 2006 and 2015. The first part of the project will involve the preparation and exploration of the data. Once correctly preprocessed, the data will be used to create a probabilistic model (a Bayesian network) for each year and then the models will be analysed and used to answer some questions. All the process followed have to be compiled into a report that will be submitted for evaluation. The project has to be developed in groups of two members.

There are many factors that may have an effect in car crashes. In this project we will explore some of them. In particular, we will use a real-life dataset that contains 684 car crashes, having for each a total of 27 variables. The dataset has been bearly preprocessed and, thus, there are still many questions that you need to consider before you can start the modelling of the data.

0.1 Preparing the data

The first problem you need tackle is the presence of missing values. There are many approaches to this problem, from the most simple (removing samples and/or variables) to the more sophisticated ones (such as imputation). One thing to bear in mind is the amount of missing information in the different variables and instances.

A second point to consider is the nature of the variables. Most of them are categorical, which is the type you will have to use for the modelling (although with the implementations in R is possible to use numerical attributes, you will have to use classical standard Bayesian networks). Note that, for converting numeric attributes to categorical, it is important to analyse how the values are distributed.

Finally, as mentioned above, one of the goals is comparing the situation in 2006 with the situation in 2015, so you will need to split the dataset into two subsets of samples.

0.2 Exploration of the data

Before going into the modelling, it is a good idea to explore the data. Some questions that may guide you are the following:

- Are all the variables informative?
- How many values do they take? How are distributed?
- What is the relationship between variables?

In this part of the project you have to analyse the variables indibidually and, if necessary, do any further processing that, according to your conclusions, you think may be good for the modelling.

Year: ICSI/KISA
Lecturer: Borja Calvo
Telephone: 943 01 50 13
E-mail: [borja.calvo]@ehu.eus

GIPUZKOAKO CAMPUSA
CAMPUS DE GUIPUZCOA
Paseo Manuel de Lardizabal, 1
20018 Donostia (Gipuzkoa)

0.3 Modelling the data

For the modelling of the data you have to use Bayesian networks. There is not a single best way to do it, so you will have to think about which the alternatives are, which make sense, and so on.

The modelling has to be done separately for the 2006 and the 2015 data, so as to compare the learned networks. Some questions to guide the comparison:

- Which variables are connected? Which are disconnected?
- Are there conditional dependencies in one year that are not in the other?
- What may be the interpretation of the differences observed?

0.4 Inference using the model(s)

Finally, we will use the two models learned to answer some questions. At least, you need to estimate the probabilities bellow, but probably there are many other questions that may be of interest. Think about (and answer) some other questions.

- What is the probability that the alcohol or drugs are involved in a crash in 2006? and in 2015?
- Does the presence of reflectors reduce the probability of a crash when the visibility is limited?
- How does the distribution of the number of cars involved change when distraction is one of the factors involved and the visbility is limited with respect to crashes where the visbility is not limited and the distraction is not a concurrent factor?

0.5 About the report

All the process, the results and the conclusions of the project explained above have to gathered into a document. The report has to be submitted in **pdf format**, and the length should not be more than 10 pages. The structure of the document should be the followings:

- **Introduction** - A brief description of the project (with your own words, do not copy-paste this document), as well as a global description of the steps taken in the analysis (preprocessing, modelling, etc).
- **Preparation of the data** - In this section you have to describe the process you have followed to prepare the data for the analysis. Also, all the decisions have to be justified.
- **Exploratory analysis** - In this section you have to explain the analysis of individual variables that you have conducted. In case this analysis has led to changes in the dataset, these also have to be described. In any case, all the decisions have to be justified.
- **Modelling the data** - In this section you have to provide the details about the process of learning Bayesian networks from data. In addition to this, the section should include a subsection where the learned networks are compared. As in the other sections, all the decisions have to be justified.
- **Inference** - In this section you have to answer the questions posed above and any other question that you may think interesting. You need to include the explanaition of the process to answer the questions.
- **Conclusions** - The last section of the report should have the conclusions drawn from the analysis. In particular, the observations in the exploratory analysis, the differences between the learned models and the answers to the questions need to be discussed.

Besides the report, the code used for the analysis have to be provided, and the readability of both the report and the code will be taken into account. As a way to improve both, you can use the concept of literate programming implemented in packages such as knitr or RMarkdown