DATASAFE: understanding Data Accidents for TrAffic SAFEty

Julyan Arbel, CR, Mistis team
Maria Laura Delle Monache, CR, Necs team
🖱 http://www.julyanarbel.com/
✉ {julyan.arbel,ml.dellemonache}@inria.fr

Stipend: 500 Euros per month.

Skills required: Tracks DS and STAT are appropriate.

Summary

This project focuses on understanding from real traffic data the behavior of traffic in the moments preceding an accident. The general approach is to use novel statistical techniques in order to learn traffic characteristics that can be used to develop new traffic models. The research team combines expertise in traffic modeling and control and Bayesian parametric and nonparametric statistics. We will use Bayesian approaches to (supervised) classification and (unsupervised) clustering in order to respectively predict collision occurrences and discover traffic patterns.

Context & Objectives

Traffic congestion is a major concern in the modern society in terms of loss of productivity, waste of time, pollution and city noise management. Two main quantities are used to describe traffic flow on a network: traffic density and average speed of cars. In the transportation literature, the graph that links the flow and the density is called fundamental diagram. The fundamental diagram is thought to be a description of the drivers' characteristic behavior and is usually assumed constant. In this project, we study how the fundamental diagram depends on time and day of the week, weather conditions, characteristics of the road and local traffic rules. In particular, we are interested in understanding if certain particular road conditions can be linked to generation of accidents.

A non-negligible part of the risk in vehicular traffic is due to individual behavior such as the driving style or the level of attention, which have a strong impact on the resulting traffic flow. It is necessary to understand if there are certain traffic situations that can cause the onset of these risks. So far, road safety has been studied mainly by means of statistical models aimed at fitting the probability distribution of the fatality rates over time or at forecasting road accidents using time series. Efforts have also been made towards the construction of safety indicators, which should allow one to classify the safety performances of different roads and to compare, on such basis, different countries. Despite this, there are no real efforts to understand if certain traffic conditions generate a risk or not. In fact, the mathematical literature offers nowadays a large variety of traffic models at all observation and representation scales, from the microscopic and kinetic to the macroscopic one. Nevertheless, there is a substantial lack of models dedicated to the joint simulation of traffic flow and safety issues. With this project, we intend to understand from data if there exists a particular link between speed and density of cars at which collisions are more likely to occur. We intend to make use of raw sensor data collected on the Rocade Sud in the framework of the Grenoble Traffic Lab using data for different days of the week and different day time. The sensor data consists of these quantities collected every 15 seconds:

- Volume: Number of vehicles that passed the sensors position in the last 15 seconds
- Occupancy: The proportion of time that the sensory was "occupied" by vehicles in the last 15 seconds.
- Speed: Average speed of the vehicles that have passed the sensor position in the last 15 seconds.

The Rocade Sud is equipped with 300 sensors along 13 km in the direction Meylan - Rondeau. Moreover, we collected accident data from the DIR (Direction Interdépartementale des Routes) Centre-Est. These data contain the time stamp and the location of the occurred accident. We will derive traffic density and flow rate from the sensor data and using state of the art statistical learning and clustering techniques we will analyze traffic behavior during periods of time that precede the accidents and we will try to derive traffic patterns. More specifically, the two tasks of collision prediction and pattern learning will be respectively addressed by classification and clustering, two of the most common tasks in statistical learning. Firstly, classification is supervised: a set of labeled data (train data) trains a learner whose further task is to distinguish and assign labels to previously unseen data (test data). Here, sensor train data is very rich: *(i) it is high-dimensional (about , consisting in three measurements (volume, occupancy, speed) over 300 sensors), (ii) it is fairly high-frequency (every 15 seconds), and (iii) it has a long track record (since 2014).* See Figure which represents speed data from one sensor over a month. Several learning algorithms will be considered, from Bayesian classification procedures, Naive Bayes to deep learning procedures, with the aim to enable prediction of collision occurrences. Secondly, clustering is unsupervised: the learner does not get any information on how the data should be sorted, but rather is expected to discover patterns in the data automatically, grouping the data by the patterns it has discovered. We are going to perform model-based clustering in a Bayesian nonparametric setting, where data sharing common sensor features will be grouped together. We are going to use Dirichlet process Gaussian mixture models, a generative model that allows us to calculate the probability of any particular set of group assignments to data points.