INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Introduction to Machine Learning — 2022/2023

Final Project

This project should be solved using Python notebooks (Jupyter) due to the ability to generate a report integrated with the code. It is assumed you are proficient with programming. All answers must be justified and the results discussed and compared to the appropriate baselines. In addition to the technical report integrated with the code, a report documenting the application of the CRISP-DM methodology should also be submitted.

Max score of the project is 4 points. The work should be done in groups (two students) or individually. In the case of groups with two members, the report should indicate an estimate of each member's contribution to the work. For example: manuel: 60%, pedro: 40%, together with a short justification. *It is mandatory to make an oral presentation and discussion of the project*.

Deadline: january 8th, 2023

Updates: · location of the data

· minor spelling (or language) corrections

The objective of this project is to apply the CRISP-DM methodology to extract knowledge from traffic accidents data, using Machine Learning methods. This is a real dataset, provided by Autoridade Nacional de Segurança Rodoviária, and it was anonymized to exclude any data that can be related to a given person or license plate.

The available data consists in three datasets containing accident data, passenger data and pedestrian data (whenever pedestrians were involved), in Portugal, over a 10 year period (2011 - 2019). The organization of the files is one file per year containing all three perspectives of that years' accidents. Notice that the data should be used exclusively for this assignment and permanently deleted after usage.

The objective of the exercise is to draw insights from this data by characterizing its subsets and also to verify the predictability of the accident types.

The project and the report should follow the phases of the CRISP-DM methodology.

To experiment with the different machine learning models, the scikit-learn[1] toolkit [Pedregosa et al., 2011] should be used.

[1] `https://scikit-learn.org/stable/index.html`

Dataset

The considered datasets correspond to accident registers, containing information on accident conditions and consequences. The data set has more that 100 000 examples, divided be year, and 28 + 30 + 43 characteristics, not all of which present relevant information. The main characteristics of the datasets are in annex.

The dataset is composed of one XLS file per year for a 10 year span. Each XLS file contains that years' partition of the three datasets. Data is available for download at:

- `https://drive.google.com/drive/folders/1meWCh8XBKh2G0tCXjCAyQZ9ty_qE8EoR?usp=share_link` (the data can only be accessed using an Iscte Google account)

Experiments

The number of accidents is very large (100 000+ examples). Experiments with this much data are difficult and long, so the first step is to sample data adequately. Sampling should be directed at your target, but should also respect the data distributions as much as possible.

Perform the following steps and report your findings:

1. Extract a random sample of 10000+ examples taking random examples from all files from 2010 to 2019. Save the sample;
2. Decide which (if any) characteristics should be normalized, discretized, or change format in any way. Perform these transformations in the sample and save the new file under a different name;
3. Verify which characteristics are seasonal (have different average values in different times of the year). Provide graphical views of these seasonal differences;
4. Verify which characteristics are unbalanced (have much more elements of one class than the others). Present graphical views of the distributions of some unbalanced characteristics you deem relevant for the problem;
5. Calculate correlations between variables and explain any high correlations found;
6. In unsupervised approaches (try at least two, k-Means and DB-Scan) after dividing the set in subsets, compare representatives or subset averages with other subsets or the general dataset to extract subset-specific characteristics (justify the choice of k in k-Means);
7. Evaluate the possibility of predicting the seriousness of the injuries based on the remaining accident data. Notice that for supervised learning the unbalanced data is serious problem, especially when the target has much more examples of one class than others.

To avoid this, collect data from all files using all cases of accidents that have other types of injuries other than "light injuries". Then sample a number of light-injuries accidents equal to the number of non-light injuries cases. Decide whether or not to use the remaining classes of injuries. In any case maintain only two classes of target values. Decide whether or not to use information from passengers and pedestrians. Report on all of these decisions. Create a reasonably balanced dataset with two classes;

8. Test three training/validation partitions in each supervised learning trial: k-fold cross-validation and two simple cross-validations with different datasets: random 30% and last 30% splits;

9. Use at least three supervised methods: Decision Tree, Multi Layer Perceptron (Neural Networks) and XGBoost. The evaluation of the generated models should include different metrics and means to better understand the errors of the supervised learning approaches, namely confusion matrices and F-scores. For any method that contains a random decision / initialization the result must be an average of 30 trials.

Explore your data with graphics, often this exploration provides critical insights. Use these graphics to help explain the conclusions drawn.

Don't forget this is a real problem, with real data, and prediction results may be disappointing. Report your findings truthfully and clearly.

References

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Data characteristics

The main characteristics of the accidents dataset are:

- accident id
- date and time
- entity in charge
- speed (local and general)
- weekday
- latitude and longitude
- number of deceased 30 days after the accident
- number of seriously injured 30 days after the accident
- number of lightly injured 30 days after the accident

- technical characteristics
- road characteristics
- district / city-hall / neighborhood / closest urban area
- type of road
- road conditions
- km of the road where accident took place
- weather
- road tracks
- intersection
- Inside / Outside urban area
- Lighting
- road markings
- type of accident
- roadwork
- obstacles
- directions
- traffic signs
- traffic-lights
- type of road-track
- trace1, 2, 3, 4
- Lane

In addition to this, the passengers dataset have extra information on the passengers' location inside the vehicle, and passenger age and gender, and the pedestrian dataset have also information in how they were involved in the accident (if any) and their resulting injuries.