Project Assignment

Machine Learning 2024 – LECD, LEEC

Informatics Engineering Department

---

1 Background

In this assignment, you must apply the methods you have learned throughout the semester to a road traffic accident severity classification problem, whose data set is available for download at Kaggle at the following link https://www.kaggle.com/datasets/saurabhshahane/road-traffic-accidents. The goal is to tackle two problems:

1) Predict the severity of an accident in a binary setting, simply distinguishing between 'Slight Injury' and the remaining classes;
2) Predict the severity of an accident, this time including all three classes, namely, 'Slight Injury', 'Serious Injury', and 'Fatal Injury'.

2 Dataset Description (adapted from Kaggle)

This dataset was collected from Addis Ababa (capital of Ethiopia) sub city police departments. It has been prepared from manual records of road traffic accidents occurring during the years 2017-2020. All sensitive information was excluded during data encoding. Two versions of the data set are available, namely, the original ('RTA Dataset.csv') and a pre-processed version ('cleaned.csv') containing the major causes (input features) for accidents. We suggest the original to be used. However, feel free to explore both. The original data set contains 32 columns and 12316 rows. The target variable (class) is 'Accident_severity', whereas the remaining variables consist of the potential input features. For more details, please check the provided link.

3 Objectives

3.1 **Scenario A (Binary Classifier)** – The objective of this scenario is to predict each instance as a 'Slight Injury' (0) or not (1).

3.2 **Scenario B (Multi-Class Problem)** – The objective of this scenario is to classify the type of injury as one of three: 'Slight Injury', 'Serious Injury' or 'Fatal Injury'.

4 Practical Assignment

4.1 Data import

Develop scripts for feature data import. Organise data into sub-sets relating to each source type you intend to test, e.g., create training, validation, and testing sets. If present, remove or find ways to handle missing data (for example, doing the mean or the median of a certain feature), check for duplicates, inconsistencies, etc. In sum, do a general data quality check before processing with the next steps.

4.2 Data Analysis

You need to understand the data you are working with. To that end, you should explore it. Use different types of data visualisation tools (histograms, pie charts, box plots, correlation plots) and analyse the data. Consider using feature selection technics and see how they affect the performance of the machine learning algorithms. Notice the class imbalance and the different features' domains in this dataset and find strategies to cope. Make sure you know your features!

Do any pre-processing, data encoding or normalisation that you deem necessary, always justifying your decisions. Do not forget to include in the final report all your findings, challenges, and steps taken.

4.2 Experimental Analysis

You should be able to design experiences in order to run the machine learning algorithms in the given data and evaluate their results. Define the appropriate performance metrics and justify your choices!

Run the experiments multiple times! You should present average results and standard deviations (of the metrics used). You may also decide to cross-validate. In the end, you should be able to choose the best classifier and evaluate them in a testing set (hold out).

Do not forget that manually inspecting the predictions of your algorithms can give you precious insights into where they might be failing (and why) and what you can do to improve them (e.g., what makes the algorithm fail in this particular case? what special characteristic does it have that makes it so hard? how can I make the algorithm better deal with those cases?). Go back and forward to the hyperparameter choice of the different machine learning algorithms until you are satisfied with

the results. It is a good idea to keep track of the evolution of the performance of your algorithm during this process. Try to show these trends in your final report to justify all the involved issues (choosing parameters, model fit, etc.). You should try to understand how they perform differently in your data.

In this project, you must use **3 different machine learning methods**: Support Vector Machines (SVM) and Neural Networks (NN) are **mandatory**, under severe grade penalty if not used. The third method is up to you to choose, except for KNN and simple Decision Trees used in the Challenge, justify the choice.

4.3 Libraries and Language

Your code must be written in Python. You are free to use Colab, Jupyter or create a script. For Neural Networks, we recommend you use Pytorch, Keras, or Scikit-learn (for more basic architectures). In any case, you are to use any Python-based framework/library.

4.4 Results and Discussion

Present and discuss the final results obtained in your Project assignment. Other authors have already studied this type of problem and related data sets. Thus, **compare** your results with the results from other sources, and **do not copy** from them.

5 Documentation

5.1 Description

Write documentation (in Portuguese or English) about your project. The documentation should include a cover page where the course name, project title, date, names, and student numbers of the authors are mentioned.

Describe the methods used for classification in such detail that the reader would be able to implement the same functions for feature extraction and classification based on your documentation and some basic background in pattern recognition. Always justify your choices, even when they are based on intuition. Do not forget to verify your assumptions! Include classification results with the given data in your documentation. At the end of your documentation, you should have a list of all references used.

5.2 Requirements, Submission, Discussion

Similar to the Challenge, this assignment is meant to be done in groups of two. If someone wants to work alone, although not advisable, it is also possible. Larger groups are not allowed.

Final Project Deadline: **May 18th, 2024.**

Deliverables:

- Exploratory Data Analysis
- Experimental design and analysis for both Scenarios, e.g., several models for the different datasets
- Final Report and Code

Project discussion: along the **week of May 20th, 2024.**