Forecasting mobile network traffic

***Abstract*—This short homework serves to assess the programming and data analysis capabilities of candidates to doctoral positions in the Networks Data Science group at IMDEA Networks Institute. The homework is aligned with the broad research interests of the group, which are at the interface between mobile networking and data science. However, it shall be regarded as a simple exercise, which is not necessarily representative of the more interesting (and complex) subjects that will be explored during the doctoral theses in the group.**

I. Tasks

The homework is about analyzing a dataset of mobile data traffic recorded in a real-world network that covers a large city. The dataset is organized in lines, each reporting traffic information for a specific combination of location and time, in the following format:

- `Square id`: the identifier of a given geographical area within the target city;
- `Time Interval`: a 10-minute time interval;
- `Internet traffic activity`: a measure of the mobile network traffic activity in `Square id` during `Time Interval`.

The dataset reports traffic information for 10,000 areas (organized in a 100×100 regular grid) during a continued period of two months. Using this dataset, the candidate is asked to complete the two tasks described next.

Task I. In the first task, the applicant has to perform a basic characterization of the data. Specifically, the following results should be produced as a minimum:

- a brief description of the hardware and software setup used to process the data;
- a plot of the probability density function of the traffic in the target city, computed over 10,000 samples that each represent the total two-month traffic in one geographical area, and with an accompanying text where the applicant comments on the result (*e.g.*, discussing the homogeneity or diversity of the distribution, and providing possible explanations for the observed behaviors);
- a figure of the time series of network traffic during the first two weeks at three areas, namely (i) the area with the highest total traffic during the two-month period, (ii) the area with `Square id` 4159, and (iii) the area with `Square id` 4556, with an accompanying text where the applicant comments on the result (*e.g.*, discussing the similarities or differences observed in the temporal dynamics of each area, and speculating on their causes).

The applicant is free (and encouraged) to provide additional results that explore any other properties of interest of the data. Supplementary figures and comments that are insightful will be considered positively during the evaluation.

Task II. In the second task, the applicant is asked to code an algorithm for one-step prediction of future traffic in a single area. Formally, let us denote as $x_a(t)$ the traffic observed at area a during time interval t. At each time t, the algorithm receives as input a history $\boldsymbol{x}_t$, *i.e.*, a vector of traffic values in past time intervals, up to t included (the exact format of the history has to be selected by the applicant, see also Section II below). The algorithm shall then produce as output an estimate $\tilde{x}_a(t+1)$ of future traffic at $t+1$ in area a. The applicant shall run the algorithm to forecast traffic in the three geographical areas identified at the second item of Task I above, during the week from December 16 to 22. The following results should be produced as a minimum:

- a self-contained description of the proposed algorithm;
- three plots reporting the superposed time series of (i) the original traffic, and (ii) the predicted traffic in the week from December 16 to 22;
- a table reporting the mean absolute error (MAE) and the mean absolute percentage error (MAPE) computed for the time series at the second item above;
- exact statistics on the training and execution time of the algorithm, with details on the process used to compute such statistics and on the hardware on which they were recorded;
- a short text where the applicant provides personal considerations on the design, performance, and possible margins for improvement of the algorithm.

The applicant can provide additional results that complement those above, only if they are strictly needed to highlight some important properties of the algorithm.

II. Methods

The reference dataset for the homework was released by Telecom Italia Mobile (TIM) for a data analysis challenge. A preliminary description of the data is provided in a paper published in Scientific Data by Barlacchi *et al.*, [1]. While the paper presents the many datasets that were made available for the challenge, only two are relevant to the homework:

- the Telecommunications activity dataset for the city of Milan (*i.e.*, data citation 5 in the paper), which contains mobile network traffic information[1];
- the Grid dataset for the city of Milan (*i.e.*, data citation 2 in the paper), which describes the tessellation of space into the areas over which such information is aggregated.

[1]As detailed in [1], the network traffic is measured in terms of number of call detail records (CDR) generated by mobile Internet sessions. This is only a proxy for the actual traffic volume in bytes, which provides however a decent estimate for the spatiotemporal dynamics of mobile data traffic.

The applicant can download both datasets from the Harvard Dataverse repository[2], at [2] and [3], respectively. In the case of the Telecommunications activity dataset, only the three fields `Square id`, `Time Interval` and `Internet traffic activity` are relevant to the homework, and all other fields can be ignored[3].

Applicants may employ any tool that they deem appropriate to solve the tasks outlined in Section I. This includes any programming language or combination of programming languages, libraries, open source code, or solutions previously proposed in the literature. For Task II, applicants have complete freedom on the choice of approach, as well as on the parametrization of the algorithm; specifically, any defintion of history $\boldsymbol{x}_t$ is allowed, (*e.g.*, in terms of the number of past traffic values it comprises, and of their associated time intervals and areas). Note, however, that the week to be predicted, from December 16 to 22, shall be used for test only, and cannot be part of the data used for training or validation, when those apply.

III. Deliverables

The applicant is required to submit the following documents by the agreed deadline:

- a Portable Document Format (PDF) document containing all text and figures detailed in Section I, which should be as concise as possible, and clearly cite all sources (libraries, open source code, web sources, scientific methods, research papers) used in the project;
- a single and well organized archive with (i) all source code, (ii) the minumum amount of input data needed to run the prediction algorithm, and (iii) precise instructions on how to run the prediction algorithm on Linux or macOS.

For any doubt or question, contact marco.fiore@imdea.org.

References

[1] Gianni Barlacchi, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri, "A multi-source dataset of urban life in the city of Milan and the Province of Trentino." Sci Data 2, 150055 (2015). https://doi.org/10.1038/sdata.2015.55

[2] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EGZHFV

[3] https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QJWLFU

[2]Note that the full data amount to around 5 GBytes, hence it is suggested to start retrieving them well in advance in case of slow Internet connections.

[3]Note that there is an error in [1] about the ordering of the fields: the country code is the third field, and all others are shifted accordingly.