Methods in Data Science

Exam Assignment

Deadline: the exam term in the student session

Overview

The purpose of this assignment is to practice the full research cycle that includes identifying an interesting question, obtaining the dataset with relevant data, cleaning the dataset, analyzing the data, and producing an informative report.

Preparation

For this assignment, I suggest you work with Jupyter Notebook (or you could use RStudio), either with Python or R. You could run your experiments on the supercomputer or on Google Colab, which one is more convenient for you.

After running your experiments, you need to produce a report. I recommend you type the report in Latex, but you could use another document editor of your choice.

Detailed instructions

In this assignment you need to fulfill several research tasks:

1. Identify one (or several) research questions of interest to a business or scientific audience
2. Obtain a dataset that can be used to address the above identified question(s). The dataset should be prepared such that to facilitate usage of machine learning methods in order to develop candidate models
3. For the selected research question(s) and dataset, you need to select several alternative ML methods and conduct scientific experiments in order to learn alternative models. Using the selected validation approach, you need to infer which is the model / method that responds better to your research question(s). You should use at least one Deep Learning method among the selected alternatives.
4. Extract valuable (business) conclusions, responding the questions extracted in the first step.

A bonus of 10 pcts will be given if you select a topic within the area of Natural Language Processing or Computer Vision.

For the tasks presented above, the focus should be on the dataset, the source code used to run the envisaged experiments and **(primarily) the report** presenting the research questions and the (business / scientific) conclusions.

Tip: Check out the *Resources* section if you need ideas for datasets you could use.

The structure of your report should be similar to a short scientific paper. In particular, it should have the following sections:

- **Introduction.** The first section should provide some context/background information, clearly identify the research question, and finally explain why the research question is important or interesting (i.e. what's the contribution), which is the (business) audience that will benefit from responding the research questions, whether the questions were approached by other previous studies, eventually using the same dataset. If your writing is concise, you can do all of this at most one page, but you can be longer if needed.
- **Data.** Several paragraphs that describe the dataset you'll be using, explain why it's appropriate for your research question, and summarize the data cleaning/wrangling you had to do to prepare the dataset for analysis. You can also present the basic characteristics of the dataset, before entering the data analysis. You can use visualization techniques, print relevant graphics and figures, and also, you can explain those figures.
- **Results and Discussion.** This is the most important part of the report. You will present in detail the experiments performed on the data. Present which analysis methods were chosen, what settings you tested for these methods, which was the validation strategy, what results you obtained for the selected methods and with the parameters tested. Compare the results obtained, and on their basis, conclude which method and final model is considered superior and how it answers the research questions described in the first part of the report. In this part of the report, you can motivate your choices by phrases such as: "I used the method *because…*", etc. This section must be a combination of text, tables and figures. It is absolutely necessary to describe and interpret the results, not just display them in tables or figures. You should also discuss the limitations of these results if you think you could obtain better results and what prevented you from doing so. You can create subsections that better structure this part of the report.
- **Conclusion** At most 2 paragraphs in which you summarize the research questions as well as the results obtained. This section should be short and concise, but should not overbid (ie draw stronger conclusions than those obtained from the analysis and justified in the previous section).

In terms of length, the entire report should be between 10 and 15 pages. Besides the report, you will need to produce the source code and the dataset. You will need to upload on Moodle and archive containing:

- The report
- The dataset. If too big, you can put a web link of the dataset
- The source code used to experiment with the data

Example questions

Here a few example research questions. You can use one of these or come up with your own research question.

- How well can we predict which customers will default on their loan?
- How well can we forecast future sales per store?
- How well can we predict success (e.g., critical acclaim, box-office success) of movies?
- Can we identify a group of people who are more receptive to a particular form of advertising?
- What are the most important predictors for employee turnover?

Submission

The deadline for the submission is **at the selected exam date in the session**. Your submission should include an archive containing three category of files:

- The report in PDF format
- The source code. If several source codes are used, please place them in a separate folder
- The dataset. If the dataset is too big, please include a web link where I can download the dataset

Pay attention!!! Reports will be verified with Turnitin for similarity. Given that a high degree of similarity is reported, you will be disqualified.

Grading

The following criteria will be used in grading your assignment. For convenience, they are grouped under the titles of the different sections of the report (10 pct are given by default).

- **Introduction (10 pct)**

– Is the reader provided with the background necessary to understand the rest of the report?

– Is the research question clearly stated?

– Is the relevance/importance of the research question adequately explained? I.e. is the contribution clear?

- **Data (15 pct)**

– Are the data appropriate for answering the research question?

– Are the data clearly described?

- **Results and Discussion (30 pct)**

– Does the analysis make sense given the research question?

– Proper methods are selected, if the experiments were run appropriately

– Are the results sensibly interpreted?

– Are results presented clearly and in an appropriate order?

– Does the section feature informative graphics and tables? Are the extracted conclusion sound motivated by the graphics, tables and the results in general?

– Are limitations and ideas for additional work specific and sensible?

- **Conclusions (5 pct)**

– Does this section provides a nice, short summary of the report?

– Are the conclusions appropriate given the analysis?

- **Source Code (30 pct)**

– Is the source code complete and can be easily re-run to reproduce everything that was done?

– How readable is the code? Is it easy to figure out what the code is doing? You can use comments to provide additional explanation of what you're doing and why.

– How efficient is the code (e.g. is there lots of duplication that could be avoided?)?

Finally, quality of writing matters. So try to make sure that your writing is clear, compelling, and concise. Please have a friend or two read the finished report and see what they find confusing.

Resources

Some websites you can use to find interesting datasets:

– Google's dataset search (https://toolbox.google.com/datasetsearch)

– Kaggle (https://www.kaggle.com/datasets)

– OpenML (https://www.openml.org)

– UCI ML (https://archive.ics.uci.edu/ml/index.php)

– KDNuggets (https://www.kdnuggets.com/datasets/index.html)

- Papers with code (https://paperswithcode.com/)