UNIL | Université de Lausanne

DATA MINING & MACHINE LEARNING FALL 2019

Instructions for term project

Prof. VLACHOS

In life we work as teams. This is what this project is supposed to teach you. If you work as a team you can achieve something that is bigger than the sum of its parts.

Working in teams can be exciting; but it can also be frustrating. There is no need to complain if someone in the group is not doing anything. If this happens (and I hope it doesn't), consider it a learning experience. And, yes, it can also happen when you work for a company or at a public institution.

For the term project your team has three options:

A) **The project option**. You will **mine actual data** for a problem of interest. These could be data from a problem from your current job, something of interest to the school, data acquired from the web, etc. You will design the data mining task, mine the data, and describe your results. You also will research existing solutions to the problem, if any have been proposed or documented. Your own data and results *need not be on par* with actual industry results; the goal is for you to get as realistic a hands-on experience as possible, given the constraints of what you have learned. A good example here, would be to use a live web service to get up-to-date data. You could also use something from kaggle or UCI, but I expect you to augment such data with other data sources.

B) **The teaching option**. You have to prepare a complete notebook where you demonstrate and **teach** some concept with text, comments links, data and Python code. Consider this as an interactive e-book. You will have to study about this topic from a few books and the Internet and then try to "teach" it via a Python notebook. The notebook should reflect the work of 4-5 weeks, so it shouldn't be trivial. Ideas here include: association rules, RFM customer value, clustering, text analytics, neural networks, etc.

C) **The research option.** You have to **implement a research paper** in Python (use it and demonstrate it in a notebook). *You can also do this alone if you want.* Consider this a starting point in case you want to do research in data mining/ML.

A) Implement this paper in Python.

B) Come to me, for implementing a (still unpublished) paper. I will provide the Matlab code. So, you should be familiar with Matlab, but the reimplementation will be easier in this case.

For option A) imagine that you are pitching your idea to a Venture Capitalist (VC).
For option B) imagine you are teaching this concept to students.
For option C) the goals are similar to B) but for recently published papers.

For all options. You will prepare

- A **Python notebook** and the related **data**,
- A **video** where you show your project (3-6mins).
- A **Github repo** where you have the related files and you collaborate with your team members. We will review the activity at github as an indication of each team member contributions.

In writing up/presenting your notebook, think of yourselves as analysts employed by or retained by a company or by a funding source (e.g., a VC firm or incubator), who wants to understand the state of the art for using data mining for the task in question. Review what has been done to date on your problem. Don't worry too much about coming up with a novel idea. It is more important to develop the idea well (within the scope of what we've discussed in class).

You should use the "data mining process" to structure your research and writeup. You should interact with me and the course assistant from the preparation of your initial ideas through your write-up, as a consulting group would interact with a firm or funding source in preparing a research report. Use your imagination, prior experience, or ask us to help to fill in any gaps between the material available and what you found already through your first Assignment.

Deliverable #1: On **October 28**, you will present me with a (initial) **idea** for your project. This should give as much detail as possible about your idea, so that I can give you feedback. Include in your proposal your ideas about: What is the exact

business problem? What is the use scenario? Is it supervised or unsupervised? How big are the data? What might be the target variable? What features would be useful? How exactly would it add business value? Etc.
You will have until **Nov 04** to **finalize the topic of your project**.

- **Between Oct 28 – Nov 04:** Setup a github workspace for the team, where you keep your current code e.g. ipynb, and any other related documents. Learn how to create a github account and learn how to use github desktop. The structure should be:

```
/code
/data
/documents (related pdfs, etc…)
Readme.md (documentation, description, link to video, etc)
```

Deliverable #2.1: On **Nov 18**, you will do your first **weekly stand-up (1min) as a team** in the beginning of the lecture. One representative from the group will explain what you have been working on and give a status report.
We will check your github to see what is there. I will check your notebook. Make sure it runs!

- By this time, you should definitely have the first data, and done some EDA.
- Potentially, have found also some more data sources to augment your initial data.

Deliverable #2.1: On **Nov 25**, you will do your second **weekly stand-up (1min)** in the beginning of the lecture.
We will check your github to see what is the status of your project. I will check your notebook. Make sure it runs!

- By this time, you should you done some initial analysis and have tried out some learning/mining techniques.

Deliverable #2.3: On **Dec 04**, you will do your final **weekly stand-up (1min)** in the beginning of the lecture.

- By this time, you should have more or less be complete with your ML/data mining modeling and have some results.
- Until Dec 12, start putting together your final results and start preparing your video.

Deliverable #3: On **Thursday Dec 12** you will submit your final report zip file containing

(a) A Python Notebook with the processing, cleaning, visualization, data mining, results, recommendations, etc. + link to **video** + link to the **github** with the code. Provide clear citations and bibliography in your notebook.
(b) The **data** (if publicly available). The data should (ideally) come from multiple sources. Try to find data to **augment** your original dataset and give more information that you can use in your analysis/visualization, etc.
(c) A **pdf** which should be the conversion of the notebook to pdf.
(d) Post the link of your video in **slack channel** of week 11.

Between Dec 12 - Dec 16: You can finetune and polish your final Powerpoint for your presentation.

Deliverable #4: In the last class meeting (**Dec 16**) you will **present** to the class the results of your research. You have up to 6 minutes for your presentation. We will allow questions from the audience.

You will get the most out of the project if you interact with me and the teaching assistant during the development of your ideas. Talk to me especially before choosing one of the business problems we cover in class (classification, regression, clustering, visualization, etc). And please feel free to come talk to me about your ideas as often as you'd like.

Structure. Your notebook write-up should include the information detailed below.

Business Understanding (take this seriously)

- Identify, define, and motivate the business problem that you are addressing.
- How (precisely) will a data mining solution address the business problem?

Data Understanding

- Identify and describe the data (and data sources) that will support data mining to address the business problem. Include those aspects of the data that we routinely talk about in class and/or in the homeworks.

Data Preparation/Pre-processing

- Specify how these data are integrated to produce the format required for data mining. *Make sure you also find additional data to augment the original features.* This is a requirement here. *(NB: data preparation can be time consuming. Get started early.)*

Data Mining/ML Algorithms used and why

- Specify the type of model(s) built and/or patterns mined.
- Discuss choices for data mining algorithm: what are alternatives, and what are the pros and cons?
- Discuss why and how this model should "solve" the business problem (i.e., improve along some dimension of interest to the firm).
- **Important: Here you are expected to experiment with more than one algorithm.** If you will do classification, then you have to try at least logistic regression, kNN and decision trees. You should play with all their parameters (eg k for knn, depth for decision trees, etc) and then report which method and parameters gave the best results. Similarly if you do regression. **Extra credit (5 points):** Use an ensemble technique to combine all these methods to give better results.

Evaluation/Graphs/Tables

- Discuss how the result of the data mining is/should be evaluated. How should a business case be developed to project expected improvement? ROI? If this is impossible/very difficult, explain why and identify any viable alternatives.
- (Bonus) Interactive tools for your notebook, so that the user can play with different parameters and see the output.

Deployment

- Discuss how the result of the data mining will be deployed.
- Discussion on what was hard to achieve, limitations
- Observations, conclusions
- Are there important ethical considerations?

Contributions

- The names of the team member, and what they did for this project. If someone contributed only for the video, or for bringing coffee (hopefully not!) that's ok. Just write that.

Resources:

- Financial Data (stocks, indexes, OECD, …): https://pandas-datareader.readthedocs.io/en/latest/index.html
- Weather Data: https://openweathermap.org/api
- List of public APIs for data: https://github.com/public-apis/public-apis#weather
- Kaggle: https://www.kaggle.com

Grading:

- (10 points): Notebook quality. Does it have clear structure? Is it easy to follow? Do you include images? Do you have interactive components? The goal of the notebook is to be an interactive and as easy to follow as possible.

- (10 points): Code quality. The different parts of the code should be included in different classes or files. Then you can call these from the notebook. Generally, for data **preprocessing** there should be one class (no need to see the cleaning in the notebook!). There should also be **another class (or python file)** for the data mining/machine learning algorithm. And another class/file for the visualization (if too complex). We will evaluate if the notebook shows good command of the language, Pandas and related libraries, and also the code abstraction.

- (10 points): Presentation. This will be based on your presentation in the last class (and your video). For this you will have to prepare 10-15 slides. We will examine, how interesting, exciting and fun is your presentation. Do you pass the message to the audience?

Some ideas for the project:

- Use twitter data to do sentiment analysis and predict whether a stock price will go up or down.
- Build a fake news detector.
- Use weather data and other publicly available data to predict/regress some value of financial/economics interest.
- Predict the IQ of a person through texts they have written.
- Create a classifier to predict if person has written a text/document or not (stylometry).

- Predict the popularity of a (newly released) song based on its lyrics.
- Predict the air quality of Lausanne (or of another location) using historical measurements.
- Predict how many tweets a user will submit in a week.
- Use the Geneva traffic accidents dataset to predict the probability of an accident to happen.
- …any other concept or idea you presented in Assignment1….