PSM Spring 2019: concluding miniproject

At the end of this course, you shall utilize what you have learnt in a little project where you carry out an explorative investigation of statistical structure in a real-world dataset. This project is defined in a rather open fashion: you can choose the dataset and the exploration methods according to your own liking. All of you are taking (or have taken last year) the companion ML course, where you have learnt a number of elementary analysis methods (especially K-means clustering and PCA) – use that knowledge in this PSM project.

The outcome of your study is a written report. Here is the assignment spelled out in as much detail as I can offer for such an open-form project:

1. Pick a dataset that you find interesting from the web. Rich sources of datasets are https://www.kaggle.com/ (a widely used platform where commercial and academic groups post their datasets for competitive crowd workouts) and https://archive.ics.uci.edu/ml/datasets.php (a repository of hundreds of benchmark datasets, some of them "good old classics", used widely in academic research).
2. *First component in your report:* Briefly motivate why you picked the dataset that you picked.
3. You may use open-access data analysis and visualization tools as you like, or program your own analysis and graphics routines. Googling "data analysis tools" or "data visualization tools" will overwhelm you with an almost unlimited choice of free toolboxes.
4. *Second component in your report:* Describe the raw dataset in mathematical correct formalism (which will often be a product of a diversity of sample spaces), and visualize the raw dataset such that the reader of your report can get a "feeling" of your data. For instance, use histograms to visualize distributions of discrete data; plot exemplary trajectories for time series data; use scatterplots for (2-dim projects of) vector data; give some illustrative text sniplets for text data or plot a few examples of image data. – When using graphics in your report, mind the advice on figures given in the "Essentials of Technical Writing" guide which you can download from the course homepage right besides this miniproject task sheet.
5. The main part of your work is to identify and document interesting structure in your dataset. You might want to look for findings like the following (incomplete list of suggestions):
    - Are there "outliers"?
    - Give an account of missing data (how severe is the missingness? In what variables? Is it in some way systematic?)
    - PCs and clusters
        - try to find intuitive descriptions of your PCs or clusters (for instance, in psychometric data one PC might be interpreted as "fear dimension")
        - PCs / clusters can be identified within the entire dataset, or subsets, or the entire dataset restricted to a subset of its dimensions
        - may require transformation from symbolic to vector data
        - for vector data: how much variance can be explained by how many PCs?
    - Often insight into relevant structure of a dataset is obtained by determining and discussing features instead of raw data variables.
    - Graphical displays of 2-dimensional continuous distributions of pairs of relevant varaibles or features that you extract

- In time series data: how "noisy" do trajectories look? Are there long-term trends or periodic ("seasonal") structures (maybe best visible after smoothing trajectories)?
- Hierarchical structures (clusters of clusters)
- Correlational information
    - visualize the cross-correlation matrix of important numerical variables
    - try to interpret this matrix! are there pairs of variables which appear to be candidates for being independent of each other? (then the 2 x 2 cross-correlation matrix restricted to these two variables should roughly be a diagonal matrix)
- for pairs of discrete variables / features, does the joint probability matrix look like it suggests independence?
- Are there linear dependencies between variables, that is, can some of your variables be well predicted by a linear combination of others? (compute linear regressions; only applicable for numerical variables)

These are all quite basic descriptors of statistical structure in your dataset. There exists a host of more advanced description tools, also nonlinear ones, which we did not introduce in the ML / PSM lectures. Some of them have been introduced in other DE courses (Data Analytics, Data Mining, Applied Dynamical Systems). While you might be adventurous enough to try some of them (may require additional self-study), this is not expected and a 100% grade can be scored without going beyond what was taught in the ML/PSM lectures.

6. *The third component in your report:* document the findings of your descriptive or analytical studies. When reporting your findings, take heed of the following guides and rules:
    - Use graphics wherever it makes sense. Humans are visual animals, and the readers of your report are humans.
    - Whatever you report, use clean mathematical formalism to express what you find. **Mastery of correct formalism is a main grading criterion for this project**. In particular:
        - Use random variable centered terminology and notation, as introduced in the PSM lecture. This does not mean that you have to use our generic symbols X or Y etc for RVs. If, for instance, a variable in your dataset is naturally called "body weight", you can call it "body weight" in your text and introduce a mnemonic variable like BW for it in formulas. But you should be aware that "body weight" or BW are RVs, and they should be used in formulas in the same ways as we used X or Y etc in the PSM course.
        - Introduce comprehensive (product) RVs and product sample spaces and distributions over these product spaces whenever it makes sense (it will often make sense).
        - Visualization toolboxes will offer you a large variety of fancy-looking graphical representations of data structures. If you use any of them, make sure that you understand the mathematical logic behind the representation that is graphically displayed, and explain it in correct formalism. I don't want to see beautiful pictures that do not come with a clear explanation of what the picture shows. If you don't have a firm grasp on the maths behind a graphics format, don't use it.
7. *Getting help*: if you have questions, ask them in class. Your classmates will be grateful.

Grading. The report will be scored on a max = 100% scale. I would expect a report of, say, at least 5 pages for a good grade, and even up to 10 pages for a perfect grade. If you consult references / tool documentation outside the ML/PSM lecture notes, give a references list. The main grading criteria and their approximate weightings are as follows:

1. Choice of an interesting and not too simple dataset, and its motivation (10%)
2. Technical correctness of your findings and formalism (50%)
3. Insight and relevance of your findings – detecting and describing structure that is useful for interpreting your dataset; insightful explanations in plain English (35%)
4. The beauty factor: clean layout, transparent structuring, high-quality graphics, good sober technical English (15%). Using Latex gives a bonus of 5% (total %score capped at 100% though).

Bonus points. Up to 10 (!) bonus points are awarded for exceptionally well done reports. What we find "exceptionally well done" is at our subjective discretion (Tianlin's and mine) and may include

- particularly rich / challenging datasets
- super-clear formulation skills (which implies super-clear understanding)
- surprising / not-easy-to-detect findings
- perfection in layout and graphics
- exceptional effort investment
- use of descriptive/analytical tools beyond ML/PSM course coverage

Bonus points are added to the final course grade undiluted. A necessary condition for bonus points is correct usage of mathematical formalism in the report.

Deliverable: the report only (no code), as pdf file named *PSMproject_<yourName>*. Tianlin will publish a place to upload your report.

Deadline: May 15, 23:59. Every day of delay costs a 10 % penalty (including bonus pts), and delay days start counting a second after midnight. Only exception from this penalty: registrar-documented illness in the days before the submission deadline.