PSM Spring 2020 Bonus Project

In this project you will carry out an explorative investigation of statistical structure in a real-world dataset, applying what we have learnt in Principles of Statistical Modeling and summarise the results in a 10 page written report.

Here is the outline of the project assignment:

1. Pick any dataset you find interesting and did not use in any other project/ course. 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:** Motivate **in your own words** why you picked this dataset. Describe the data acquisition. Which steps did you have to take to download the data.

3. **Second component in your report***:* Describe the raw dataset in mathematical correct formalism (which is typically a product of several different sample spaces), define the universe in which the data has been taken, what are the RV functions, and the data value spaces S.

4. Describe whether you had to clean the data, are there any missing values, how did you deal with missing values? Show excerpts of the raw data.

5. **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. Use heat maps to visualise time series data, which often reveals correlations and seasonalities. Use scatter plots for vector data. Give some illustrative text excerpts for text data. Plot 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 find also on Moodle.

6. **Third component in your report***:* 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?)
- Principle Components (PCs) and clusters
    - 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.
- Define transformations if you had to apply them to your DVS.
- Graphical displays of 2-dimensional continuous distributions of pairs of relevant variables 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
    - Calculate the Pearson correlation coefficient.
    - Define the cross-correlation matrix and calculate it for important numerical variables, visualise it using f.e. a heatmap.
    - Interpret this matrix: are there pairs of variables which are strongly correlated, or anticorrelated or appear to be candidates for being independent of each other?

- 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 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 PSM lectures. Some of them have been introduced in other DE courses (Data Analytics, Data Mining, Applied Dynamical Systems). You may of course apply them, but this is not expected and a 100% grade can be scored without going beyond what was taught in the PSM lectures. For any methods you employ be sure to document them properly.

7. **Fourth component in your report**: document the findings of your Data analysis and modeling. When reporting your findings, please take into account the following guides and rules:

- Use graphics wherever it makes sense. Humans are visual animals, and the readers of your report are humans. Give all information needed to understand the figures in the figure captions, including a reference to the source it was taken from (if it is not your own figure). Discuss the figures then in the mean text, refering to it by the Fig. number.
- 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. It is often convenient to use our generic symbols X or Y etc for RVs. But, for instance, a variable in your dataset is naturally called "body weight", you can call it "BW" in your text and 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. Especially label the axis in the figure, and define it in the figure caption. Beautiful pictures without clear explanation of what the picture shows will count negatively. If you don't have a firm grasp on the maths behind a graphics format, don't use it.

7. If you have questions, ask them during the lecture or send me an email.

Grading. The report will be scored on a max = 100% scale with typically 10 pages, including references. Be sure to cite all references and tool documentations you used. The main grading criteria and their approximate weightings are as follows:

1. Choice of an interesting and not too simple dataset, and its motivation (For that the data size does not have to be Big, several thousand data points are often sufficient) (20%).

2. Technical correctness of your findings and formalism (40%).

3. Insight and relevance of your findings – detecting and describing structure that is useful for interpreting your dataset; insightful explanations in plain English and correct technical terms. (20%) .

4. The form: layout, transparent structuring, high-quality graphics, readable labels in diagrams, complete citations in text and of figures, good sober technical English (20%).

Any of the following may influence the grading positively:

- Particularly rich / challenging datasets

- Super clear formulation (which implies super-clear understanding)

- Interesting, not obvious findings, carefully documented and explained with data analysis.
- Perfection in layout, graphics
- Careful literature, source documentation
- Clear Usage of descriptive and analytical tools introduced in the PSM course.
- **Deliverable: Upload** the report with code as pdf file named *PSMproject_<yourName>* **on Moodle** .
- **Deadline:** May 31. Every day of delay after that costs a 10 % penalty (exception of course: medical excuse reported to registrar and email to me before the submission deadline).