7PAM2000 Applied Data Science 1

Assignment 3: Clustering and fitting

This third assignment will focus on clustering and fitting. This time you are expected to produce a poster which could be used for a poster presentation.

We will again look at exploring public data from the World Bank, and specifically country-by-country indicators related to climate change: `https://data.worldbank.org/topic/climate-change`. You may find additional relevant indicators (e.g. GDP per capita) using the complete list `https://data.worldbank.org/indicator`. Note that not all countries have entries for the most recent year(s).

Your goal is to:

- Find interesting clusters of data. Note that for meaningful clusters it is often a good idea to look at normalised values like GDP per capita, CO_2 production per head, CO_2 per $ of GDP or fraction of a sector. You might look at most recent values or compare recent values with, say, values 30 or 40 years ago or use total historic values.

  Use at least one of the clustering methods from the lecture. Clustering works best when the data are normalised (see Practical 8). Note that you usually want to show the original (not normalised values) to display the clustering results. One way to achieve this is to add the classifications as a new column to the dataframes and use logical slicing. Produce a plot showing cluster membership and cluster centres using pyplot.

- Create simple model(s) fitting data sets with `curve_fit`. This could be fits of time series, but also, say, one attribute as a function of another. Keep the model simple (e.g., exponential growth, logistic function, low order polynomials). Use the model for predictions, e.g. values in ten or twenty years time including confidence ranges.

  Use the attached function `err_ranges` to estimate lower and upper limits of the confidence range and produce a plot showing the best fitting function and the confidence range.

- You do not need to use the same data sets for clustering and fitting, but one approach could be: find clusters of countries, pick one country from each cluster and compare countries from one cluster and find similarities and differences, compare countries from different clusters or pick a few countries from one clusterand compare with other regions.

- You do not need to focus on CO_2 or climate change. The choice of topic is yours.

Content and presentation of the report

Again, you are expected to use your initiative and "tell a story" with the data. You should use appropriate visualisation and provide a text narrative to communicate and explain your findings.

There a some differences between a good report and a good poster. Text in a good poster is shorter. Sometimes use of itemisation or similar could be a good idea. Some information, like comparison between countries, is often better presented in an overview table than in a detailed text.

Not every detail should be explained. A report must contain all necessary information. Readers can ask in a poster session. Still the text should allow a basic understanding without further information. Self-contained graphs are an important tool. Similar to a report good posters often combine information from graphs to draw conclusions or follow up on insights/questions from one graph with another graph.

Your poster should include an introduction and explicit conclusions summarising the important results. Do not include technical information. Your audience is interested in the results not how you got there. Very(!) short remarks on methods used can sometimes we appropriate, but not more.

Format guidelines for the poster

Produce a poster presenting your results. Due to time constraints we will not do poster presentations, but when designing the poster have in mind you doing a poster presentation to an audience of two or three.

- *Visualisation is even more important for a poster than for a report..* A good graph can tell more than one page of text. Make the graph self-contained. Provide information needed to understand it in the graph: meaningful labels, legend and – possibly – a good title. This is more important for a poster than in a report where it is easier for the reader to switch between text and graph.

- The choice of software is your decision, but PowerPoint is often used. Example posters are available in the example posters unit. A PDF file with the poster should be uploaded.

- The poster should be size DIN A0 (841 mm × 1188 mm). To enter the size open the *Design* tab, open the *Size* dialog and *Customise*. The smallest font size should be 20. (Scaled) fonts in the graphs can be smaller but should still be legible from a distance.

- You can divide a page into columns: go to the `Layout` tab, select `Columns` and the number of columns.

- You can use portrait or landscape format. In the `Layout` tab select `Orientation`.

- The design of a good poster is different from a good report. A good optical structure is important. The optical structure should guide the reader through the poster. *This can be achieved in various ways.* Examples are use of colour, arrangement of the text, highlighting text, positioning of plots.

  Avoid long paragraphs of unstructured text. Sometimes an itemised list is a good way to structure information. Or consider an overview table comparing, say, countries. *Keep in mind that more than one person may be reading the poster from a distance at the same time.*

Criteria for the coding quality mark.

- Adherence to the PEP-8 guidelines and the good style guide.

- Well structured and commented program following the good style guide, good use of functions (see mark sheet). No spaghetti code please.

- Good use of your repository with an appropriate level of commitments.

What data can I use?

Your report needs to use Worldbank data. Additional files can be used (e.g. sales of electric vehicles not included in the Worldbank data). You can make use of Worldbank data not included in the collection, of course.

What modules can I use?

- Use of one of the clustering methods from the lecture and `curve_fit` is expected (see mark sheet). Use pyplot or plot methods of pandas if appropriate.
- Functions from other modules can be used in addition. The lecture material provides sufficient tools for the assignments, but sometimes you'll find functions in other modules doing something special.

What to submit?

- PDF file of the poster. PDFs are preferred because they avoid potential format problems (different defaults on different systems). The poster should be uploaded as is and not be wrapped into a zip file or similar.
- Your program as python file. Notebooks are depreciated.
- **Do not** upload data files. Worldbank files do not need to be referenced. Links to other data suffice.
- *Repositories:* Repositories are mainly to store version of programming code. We expect repeat commits for full marks. Other material can also be committed, but that is optional. Post your github link as a comment or upload a zip file with your local repository.