7PAM2000 Applied Data Science 1

Assignment 3

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). Write a function to do the normalisation. Note that you usually want to show the original (not normalised values) to display the clustering results.

- 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 years time including confidence ranges.

  Visualisation is important. A good plot can tell more than one page of text. Provide information needed to understand the plot in the plot: 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.

  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. Create a function (or functions) which does the fitting and plotting.

- 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 do comparative fitting or pick a few countries from one cluster and find similarities and differences.

- Produce a poster presenting your results. Due to time constraints we will not do poster presentations and vivas, but when designing the poster have in mind a poster pinned to the wall.

  The poster should be size DIN A1 (60 cm × 80 cm) . The format can be vertical or horizontal. To enter the size open the *Design* tab, open the *Size* dialog and *Customise*. The smallest font size should be 20. *Keep in mind that more than one person may be reading the poster from a distance at the same time.* The choice of software is your decision, but PowerPoint is often used. Example posters are available in the example posters unit.

The design of a good poster is different from a good report. A good optical structure is important. *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.

Criteria for the coding quality mark.

- Adherence to the PEP-8 guidelines and the style guide..
- Well structured and commented program, good use of functions. No spaghetti code please.
- Good use of your repository with an appropriate level of commitments.