DPRPy 2024/2025

Homework assignment no. 3 (max. = 30 p.)

Maximum grade: 30 p.

Homework should be sent via the `MS Teams` platform. You should include all files containing solutions to tasks as one `.zip` archive named i.e. `Last-name_First-name_HA_3.zip`.

Note: Your solutions may be in one or multiple Jupyter Notebook files, knitr + Markdown files, `.py` or `.R` scripts, etc. Just make sure that each file is clearly named.

1 Tasks description

1.1 Part 1

In this task, we will use data provided by the Stack Exchange network https://archive.org/details/stackexchange. However, this time you can choose the forum topic. Let us recall that for each forum available data consists of few `xml` files, e.g.:

- Posts.xml
- Users.xml
- Comments.xml
- Votes.xml
- Tags.xml
- Badges.xml
- PostLinks.xml
- etc.

1.1.1 Task 1 [10 p.]

Based on the data from the selected forum, in programming language of your choice (`R` or `Python`) prepare the following charts:

1. Histogram;
2. Box-and-whiskers plot;
3. Bar chart;
4. Heatmap;
5. Line or bubble chart;

Make sure that resulting plots are aesthetically pleasing and clear to read. To each chart add a caption / brief description of what it represents.

For example, you may consider the posts score and prepare the histogram with density function estimation added. You may then plot the box-and-whiskers plots of score when divided by the post type, etc.

1.1.2 Task 2 [10 p.]

Use ChatGPT to prepare any chart you want based on the data corresponding to your selected forum (i.e. choose a topic that you find interesting). Then, based on the code suggested by ChatGPT (you can choose whether it should be in R or Python), create the aforementioned plot. The main part of this task is to analyze the solution provided by ChatGPT. Describe what the chart presents, its strengths and weaknesses, whether it is readable, what could be improved, etc. In in your submission include following:

1. You query to chatGPT;
2. Code generated by ChatGPT and figure created with it;
3. Your analysis of obtained solution.

Please remember: the more interesting figure, the better!

1.2 Part 2

1.2.1 Task 3 [10 p.]

Consider algorithms that aim in finding the grouping of a data set that reflects the underlying structure of said data. Now, assume that you have results of a few, selected algorithms of this kind on multiple data sets. The goal of this task is to prepare the chart that will allow the reader to easily compare quality of results obtained by each algorithms across benchmark data sets.

The detailed instructions on how to generate the data to plot are included in attached `.R` and `.py` files. You may use the programming language you want, i.e. `R` or Python.

Important

Github repositories needed:

The "correct" labeling, i.e. ground truth: https://github.com/gagolews/clustering-data-v1

Note that for each data set the file with reference labelling is named in following name:

`collection_name/data_name.labels0.gz`,

e.g. `fcps/atom.labels0.gz`, where `collection_name == fcsp` denotes the source of data and `data_name == atom` is the name of the data.

The results generated by algorithms: https://github.com/gagolews/clustering-results-v1/

Note that files with results are named with the convention:

`original/algorithm/collaction_name/data_name.resultK.gz`,

where:

- `original` simply mean that no manipulation of data were performed;
- `algorithm` denote the method that was used to generate results;
- `collection_name`, as previously, denotes the data source;
- `data_name` indicate for which data set the results were calculated;
- `K` denotes the number of groups that were determined.

For example,

`original/fastcluster_average/fcps/atom.results2.gz` means that the method used was fasctcluster_average algorithm, the results were calculate for atom data sets from fcps collection and two groups were find.

In order to evaluate how good the quality of results is we must measure its agreement with reference labelling. We will do that with the usage of Adjusted Rand Index (ARI, in `R` - `mclust::adjustedRandIndex()`). In general, the better the results the closer ARI value is to 1. Values around 0 indicate the very bad quality of results.

R script that allow to generate the assessment of each method on each data set is given in `Homework_Assignment_np_3.R` files included in Teams.