CASE STUDY

Tasks for Course:

DLBDSMLUSL01 – Machine Learning – Unsupervised Learning and Feature Engineering

CONTENT

1. TASKS

In this section, you can select one of the listed case studies to work on (see sections 1.1, 1.2, 1.3).

When working on your case study, please consider the task described in the respective case study itself.

Note on copyright and plagiarism:

Please take note that IU Internationale Hochschule GmbH holds the copyright to the examination tasks. We expressly object to the publication of tasks on third-party platforms. In the event of a violation, IU Internationale Hochschule is entitled to injunctive relief. We would like to point out that every submitted written assignment is checked using a plagiarism software. We therefore suggest not to share solutions under any circumstances, as this may give rise to the suspicion of plagiarism.

1.1 Task 1: Mental Health in Technology-related Jobs

You are working at a technology-oriented company. The Human Resources (HR) department is about to start a pre-emptive program towards mitigation of mental health issues amongst the company’s staff. HR has contacted you as one of the company's data scientists to support the program with quantitative analyses. You are provided with the results of a survey, which has been conducted amongst technology-oriented employees and which is considered representative for your company as well. The challenge in working with this data set lies in its high dimensionality and complexity so that it is not straightforward to interpret and transfer the results to the structures of your organization. Another challenge in working with this data is missing values and non-standardized textual inputs. HR has asked you to provide a better overview over the data in ways which are more easily interpretable and transferable to improvements at your workplace. The goal is to categorize participants of the survey according to their answers and to provide visualizations supporting the interpretation of these clusters. These visualizations should convey a perspective onto the data set, which is reduced in complexity and dimensionality and yet preserving the main characteristics of the whole available data. You should also provide insight into each individual identified cluster of participants and their main characteristics. Ultimately, this will help to identify potential points of leverage for the planned program, which will be addressed with targeted measures.

Your written documentation and concept for implementation will be evaluated. The accompanying code for your approach to this use case should be downloadable from the web and a link should be provided as part of the use case documentation.

Note: There is not *one perfect* solution. It is important to keep the concepts and techniques you have learned during the course in mind and to carefully prepare and process the data accordingly. At the end of each step of this use case, you should critically assess whether all necessary operations have been conducted and argue towards the decisions you have made during this process.

Tips: Start by exploring the available data. Some descriptive statistics and explorative visualization usually make sense at this point. After you have understood the available data, design a plan for your approach to this use case. This plan has not to be perfect as it will change when you iteratively proceed. You will probably have to spend quite some time with pre-processing the data to build towards a data structure which can be processed by the respective machine learning libraries. You might also find that feature engineering can improve your data to better capture key information in a way that it is usable by machine learning algorithms. When it comes to the actual implementation, keep it simple in the first iteration and try to come up with quick solutions. Building from this, you can elaborate and take some steps back to improve the quality of your work.

Data: For this use case, imagine that HR has provided you with the results of the conducted survey. You can choose to find a suitable, freely available dataset on your own or you can use the data which is available from the following webpage: https://www.kaggle.com/osmi/mental-health-in-tech-2016

Alternatively, the data can also be downloaded using the following link: https://iubhfs-my.sharepoint.com/:f:/g/personal/c_mueller-kett_iubh-fernstudium_de/EhnpanaFz7pDnwl78sEWpmkBAUo5NgrIDBmTWUqjM0m1rQ?e=5bVs1n

1.2 Task 2: Policing Equity

There has been a lot of concerns about policing equity in your local community. Driven by the political discourse, your local municipal administration has decided to quantitatively investigate this issue and has contacted you as a freelancing data scientist to support this effort. For a couple of years, there has been a standardized data collection process with respect to policing activities in place. Until the present day, the dataset has grown to a considerable size and overlooking patterns in the data has become an intricate task. In the first step of this use case, you investigate the provided dataset towards homogeneous categories of similar policing incidents. You will find that the geographic information is contained in the dataset and might be useful for the investigation of patterns within the data. The goal is to get a better overview over policing activities in your community which should, ultimately, lift the political discussion to a more informed level. Your goal in this use case is to reduce the complexity of the dataset which is, at first, hard to overlook and to provide visualizations which capture the main characteristics of the whole dataset. For this purpose, you consider different techniques for dimensionality reduction. You should also provide insights into preferably homogeneous clusters of policing activities and visualizations, which allow to interpret these clusters. Finally, descriptive statistics about each group along with the number of incidents per cluster should be provided.

Note: There is not *one perfect* solution. It is important to keep the concepts and techniques you have learned during the course in mind and to carefully prepare and process the data accordingly. At the end of each step of this use case, you should critically assess whether all necessary operations have been conducted and argue towards the decisions you have made during this process.

Tips: Start by exploring the available data. Some descriptive statistics and explorative visualization usually make sense at this point. After you have understood the available data, design a plan for your approach to this use case. This plan has not to be perfect as it will change when you iteratively proceed. You will probably have to spend quite some time with pre-processing the data to build towards a data structure which can be processed by the respective machine learning libraries. You might also find that feature engineering can improve your data to better capture key information in a way that it is usable by machine learning algorithms. When it comes to the actual implementation, keep it simple in the first iteration and try to come up with quick solutions. Building from this, you can elaborate and take some steps back to improve the quality of your work.

Data: For this use case, imagine that your local municipal administration has provided you with the data on policing activities. You can choose to find a suitable, freely available dataset on your own or you can use the data which is available from the following webpage: https://www.kaggle.com/center-for-policing-equity/data-science-for-good?select=ACS_variable_descriptions.csv

Alternatively, the data can also be downloaded using the following link: https://iubhfs-my.sharepoint.com/:f:/g/personal/c_mueller-kett_iubh-fernstudium_de/El-dI09qEYhFsApCJauOdb8BdaPFXneP8xYbFIbJqFSUug?e=Vw4ghJ

1.3 Task 3: Categorizing Trends in Science

You are working as a data scientist at a company which has defined the strategic objective to position itself more towards research and academic cooperation. During an early stage of this effort, the company wants to investigate what the current topics in science are and for which topics an advanced academic corporation could make sense. For this purpose, you are engaged with no simpler task than to provide a comprehensive quantitative overview on current topics in science. You have access to a large archive on scientific papers which listing a plethora of recently published scientific articles, which is hard to overlook due to its sheer volume. The goal of this use case is to provide insights into current trends in science based on preferably homogeneous clusters and to provide views which are reduced in dimensionality, but still capture the main traits of the provided dataset. As the data contains, in parts, text which must be analysed, the containing information will need pre-processing accordingly.

Note: There is not *one perfect* solution. It is important to keep the concepts and techniques you have learned during the course in mind and to carefully prepare and process the data accordingly. At the end of each step of this use case, you should critically assess whether all necessary operations have been conducted and argue towards the decisions you have made during this process.

Tips: Start by exploring the available data. Some descriptive statistics and explorative visualization usually make sense at this point. After you have understood the available data, design a plan for your approach to this use case. This plan has not to be perfect as it will change when you iteratively proceed. You will probably have to spend quite some time with pre-processing the data to build towards a data structure which can be processed by the respective machine learning libraries. You might also find that feature engineering can improve your data to better capture key information in a way that it is usable by machine learning algorithms. When it comes to the actual implementation, keep it simple in the first iteration and try to come up with quick solutions. Building from this, you can elaborate and take some steps back to improve the quality of your work.

Data: For this use case, imagine that you have been given access to an archive of recently published scientific works. You can choose to find a suitable, freely available dataset on your own or you can use the data which is available from the following webpage: https://www.kaggle.com/Cornell-University/arxiv

Alternatively, the data can also be downloaded using the following link: https://iubhfs-my.sharepoint.com/:f:/g/personal/c_mueller-kett_iubh-fernstudium_de/ErBgsM5LevNCmSyuTXlZv28B6g49YQyZ6j_I8gmqdLyt5w?e=c9OjgJ

2. ADDITIONAL INFORMATION FOR THE EVALUATION OF THE CASE STUDY

When conceptualizing and writing the case study, the evaluation criteria and explanations given in the writing guidelines should be considered.

3. TUTORIAL SUPPORT

Students have the option to make use of any one of several opportunities to get support for their case study analysis with the course tutor. Taking advantage of these opportunities is the responsibility of the student and the use of these services is voluntary. It is possible to contact the tutor regarding formal and general questions about working on the case study. Please note: a review of outlines and aspects of the presentation is not intended here, since the student’s ability to work independently is part of the evaluation and counts as a part of the overall assessment. There are however general tips for developing the case study to help you getting started.