School of Computing & Communications | Lancaster University

SCC403 – Data Mining

Coursework Assignment

1 Introduction

The objective of the assignment is to conduct data analysis on a real life data set, which concerns climate. This includes selection and justification of the specific methods for data pre-processing, clustering and classification, their implementation and analysis of the results and well annotated code. It is expected that your analysis should include:

- data pre-processing (to achieve top marks a well justified variety of specific pre-processing techniques is expected)
- clustering (to achieve top marks a well justified variety of specific clustering techniques is expected)
- data classification (to achieve top marks a well justified variety of specific classification techniques is expected)

All choices must be justified through analysis and comparison. Analysis and understanding of the methods, algorithms and the overall process are the most important elements in addition to the implementation skills (code, presentation) and the results. You are expected to critically analyse the results of applying these techniques, and demonstrate a clear understanding of the purpose and processes of data analysis.

In addition to your report, please submit your source code, including comments, plots, and an analysis of the results. *You are free to use libraries, but students that implement methods from scratch will be more likely to receive higher grades.*

We recommend using Python, since it is the language that we are using in the labs, but you are free to use a different languages if you prefer. Note, however, that we may need to contact you for clarification, if we do not understand your code, or if we believe that your code is not running correctly.

2 Data-Set

You are expected to use the set of climate data provided in the file $'SCC403CWClimateData.csv'$. This data is a subset of publicly available (from `https://www.meteoblue.com/`) data about climate in Basel, Switzerland which contains 1763 records (18 features, a 18 dimensional vector) of data from the summer and the winter seasons of the period from 2010 to 2019 period. The meaning of each column of data is listed below:

- Temperature (Min) ^oC.
- Temperature (Max) ^oC.
- Temperature (Mean) ^oC.

- Relative Humidity (Min) %.
- Relative Humidity (Max) %.
- Relative Humidity (Mean) %.
- Sea Level Pressure (Min) hPa.
- Sea Level Pressure (Max) hPa.
- Sea Level Pressure (Mean) hPa.
- Precipitation Total mm.
- Snowfall Amount cm.
- Sunshine Duration min.
- Wind Gust (Min) Km/h.
- Wind Gust (Max) Km/h.
- Wind Gust (Mean) Km/h.
- Wind Speed (Min) Km/h.
- Wind Speed (Max) Km/h.
- Wind Speed (Mean) Km/h

The lectures and tutorials will provide you with the necessary tools to conduct your analysis. You may also include additional analysis methods that you have researched separately,that may help derive your conclusion (this is not compulsory).

It is expected that your analysis will include;

1. pre-processing,
2. clustering, and
3. classification of the data.

For the classification task you will need to provide suitable and logical labels to different classes of data, for example, "cold, dry, windy", "wet and windy", "dry, warm with no wind", etc. You are expected to critically analyse the results of applying these techniques, and demonstrate a clear understanding of the purpose and processes of data analysis.

The deadline for submission is: 4pm, 11 December 2020, Friday. The cut-off deadline is 4pm, 14 December 2020, Monday (with late submission penalty incurred which is 1 letter grade or 10%). Submissions after this deadline cannot be accepted according to the University regulations.

In case your code is unclear to us you may be contacted for interview. If you fail to reply or attend the interview your code could be marked as "not working". Since the deadline is in the end of the Michaelmas term, it is acceptable to attend the interview by teleconference in case you are not in Lancaster.

3 Marking Scheme

For the overall report marks are allocated as follows:

- Structure and presentation (5%)
- Language and style (6%)
- Level of understanding (12%)
- Depth of analysis (12%)
- Use of literature and references (5%)

Each of the three parts (pre-processing, clustering, and classification) will form 20% of the total mark split as follows:

- Working, well annotated code and results (8%)
- Justification of selected methods (6%)
- Independent research and use of methods not given in the lectures (6%)

At the end of this document there is an Appendix, which explains what a mark means in Lancaster University and includes suggestions for a well-written report.

The length of the report should not exceed 6 pages. You can use double column format, e.g. the so-called IEEE style as described in the Appendix. You may include an Appendix (2 pages maximum) after the main report (8 pages in a total).

4 Tasks description

4.1 Pre-processing

Pre-processing includes data standardisation and/or normalisation, detecting and removing anomalies, missing values (if any), feature selection and/or extraction (the latter are optional). Pre-processing provides an insight into the data correlations and patterns.

If you choose to use the Principle Component Analysis (PCA) method, you can extract new, orthogonal (independent) features, which are a linear combination of the original ones (which carry a clear physical meaning, such as temperature or pressure). If you choose to use PCA, please, comment on the amount of variance, interpretability and the link with the original features. You should also plot the results using, for example, the one or two of the principle components which contain most of the variance.

4.2 Clustering

Choose at least two clustering algorithms. To achieve top marks one of the methods should be from independent research.

Develop the programme and explain the functionality of the algorithms in as much detail as you can. Compare the results and limitations of each of the algorithms that you have used.

4.3 Classification

Climate data set does not have specific class labels - you may label the clusters you obtained in the previous Task with some meaningful names, e.g. *"Cold windy days"*, *"Hot dry days"*. Alternatively, you may choose to use numbers 1,2,... or letters.

The task is to train at least two classifiers of your choice on a part of the data (again, you may choose how to split the data), perform cross-validation, evaluate the performance of the classifiers and report this.

When analysing the performance of the classifiers you should use precision/recall, F1 score and classification accuracy as well as computational time for training the classifier as a measure of complexity.

5 Additional Comments

You must report in an “acknowledgements” section the use of any libraries, readily available online code, and code from online tutorials. Additionally, you are free to discuss your work with colleagues, but you must also report in the “acknowledgments” section if anyone has helped you significantly. Remember that using others’ work without giving the due credit is an act of *plagiarism*, and it is not a good academic practice.

APPENDIX

Example of the style of the report

Title of the Report

Subtitle as needed

Author's names, Student number
line 1: dept. name of organization
line 2-name of the programme and module

Abstract— **Briefly describe the outline of your report.**

I. Introduction

Here you have to provide the background review. of the existing approaches stressing the ones that have been actually used. Critically analyse and compare alternative techniques and methods. Try to go beyond what was given in the lectures using external sources and references.

II. Pre-processing

Here you have to provide a description and description and the results of pre-processing techniques that are relevant and stress those that you actually used in your work. Provide the software code that you used to obtain the results in an Appendix. Do not forget to justify your choice.

III. Clustering

Here you have to provide a description and the results of clustering techniques that are relevant and stress those that you actually used in your work. Provide the software code that you used to obtain the results in an Appendix. A very important part of your report is the critical analysis of the results. Why have you used the stated methods? What are the advantages and limitations of the algorithms that you have used?

IV. Classification

Here you have to provide a description and the results of classification techniques that are relevant and stress those that you actually used in your work. Provide the software code that you used to obtain the results in an Appendix. A very important part of your report is the critical analysis of the results. Why have you used the stated methods? What are the advantages and limitations of the algorithms that you have used?

V. Conclusion

Describe briefly what has been done, with a summary of the main results. Discuss here possible future developments (what you would have done more). What is distinctive about the results you have obtained?

VI. References

The template will number citations consecutively within brackets [1]. The sentence punctuation follows the bracket [2]. Refer simply to the reference number, as in [3]—do not use "Ref. [3]" or "reference [3]" except at the beginning of a sentence: "Reference [3] was the first ..."

Number footnotes separately in superscripts. Place the actual footnote at the bottom of the column in which it was cited. Do not put footnotes in the reference list. Use letters for table footnotes.

[1] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001

[2] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* Heidelberg, Germany: Springer Verlag, 2001

[3] Angelov, P.: Autonomous Learning Systems: From Data Streams to Knowledge in Real Time. John Wiley and Sons (2012).

[4] Angelov, P.: Outside The Box:An Alternative Data Analytics Frame-Work. *Journal of Automation, Mobile Robotics & Intelligent Systems*. Vol. 8, 29–35.

Appendix

Please include here additional experimental results or additional details.

Presenting someone else's work as your own in an assignment without proper citation of the source is an act of **plagiarism**. More information about Lancaster University Plagiarism Framework can be found at https://www.lancaster.ac.uk/academic-standards-and-quality/information-and-resources/policies-and-

What a Mark Means in Lancaster University

70 + (Distinction)

Critical Understanding of Topic

Excellent understanding and exposition of relevant issues; insightful and well informed, clear evidence of independent thought; good awareness of nuances and complexities; appropriate use of theory.

Structure of Research

Substantial evidence of well implemented independent research and / or Substantial evidence of well selected evidence to support argument.

Use of Literature

Excellent use of literature to support argument /points.

Conclusion

Excellent; clear implications for theory and/or practice.

Language

Excellent; a delight to read.

Structure and Presentation

Arguments clearly structured and logically developed; sensible weighting of parts; meaningful diagrams; properly formatted references.

65 – 69% (Very Good Pass)

Critical Understanding of Topic

Clear awareness and exposition of relevant issues; some awareness of nuances and complexities but tendency to simplify matters; based on appropriate choice and use of theory.

Structure of Research

Some evidence of independent research reasonably well implemented and / or some evidence of identification of suitable evidence to support argument.

Use of Literature

Good use of literature to support arguments.

Conclusion

Very good; draws together main points; some implications for theory and/or practice

Language

Carefully written; negligible errors.

Structure and Presentation

Arguments clearly structured and logically developed; good weighting of parts; meaningful diagrams; properly formatted references.

60 – 65% (Good Pass)

Critical Understanding of Topic

Shows awareness of issues and theories; attempts at analysis but tendency to lapse into description

Structure of Research

Some evidence of independent research reasonably well implemented and / or some evidence of identification of suitable evidence to support argument.

Use of Literature

Use of standard literature to support arguments.

Conclusion

Reasonable conclusion that summarises essay; a few implications for theory and/or practice.

Language

A few errors; generally satisfactory.

Structure and Presentation

Arguments reasonably clear but undeveloped; some meaningless diagrams or poor structure.

50 – 59% (Pass)

Critical Understanding of Topic

Work shows understanding of topic but at superficial level; no more than expected from attendance at lectures; some irrelevant material; too descriptive.

Structure of Research

Insufficient evidence of independent research and / or very limited evidence used to support argument.

Use of Literature

Use of secondary literature to support arguments.

Conclusion

Conclusion does not do justice to body of essay; too short; no implications.

Language

Some errors; grammar and syntax need attention.

Structure and Presentation

Arguments not very clear; poor organisation of material; poor use of diagrams; poor referencing.

45 – 49% (Marginal Fail)

Critical Understanding of Topic

Establishes a few relevant points but superficial and confused; much irrelevant material; very little or no understanding of the issues raised by the topic or topic misunderstood; content largely irrelevant; no choice or use of theory; essay almost wholly descriptive; no grasp of analysis with many errors and/or omissions.

Structure of Research

No evidence of independent research and / or No attempt to identify suitable evidence to support argument.

Use of Literature

Relies on a superficial repeat of class notes.

Conclusion

No recognisable conclusion.

Language

Frequent errors; needs urgent attention.

Structure and Presentation

Arguments often confused and undeveloped; no logical structure; very poor organisation of material; many meaningless diagrams; negligible referencing.

0 – 44% (Clear Fail)

Critical Understanding of Topic

Establishes a few relevant points but superficial and confused; much irrelevant material; very little or no understanding of the issues raised by the topic or topic misunderstood; content largely irrelevant; no choice or use of theory; essay almost wholly descriptive; no grasp of analysis with many errors and/or omissions.

Structure of Research

No evidence of independent research and / or No attempt to identify suitable evidence to support argument.

Use of Literature

No significant reference to literature.

Conclusion

No recognisable conclusion.

Language

Frequent errors; needs urgent attention.

Structure and Presentation

Arguments often confused and undeveloped; no logical structure; very poor organisation of material; many meaningless diagrams; negligible referencing.