School of Computing & Communications | Lancaster University

SCC403 – Data Mining

Plamen Angelov, Leandro Marcolino

Final Coursework

1 Introduction

A data scientist must be able to load data-sets, pre-process the data, and apply various algorithms for analysis. In particular, data partitioning (e.g., clustering), and classification are very common approaches for handling complex data-sets. We studied several techniques in the lecturers and lab sessions, and in this coursework you will exercise these techniques on three different data-sets. Additionally, you can learn even more by going beyond the algorithms that were discussed in class.

Therefore, it is expected that your analysis will concern:

- data pre-processing, including normalisation, standardisation, dealing with class imbalance, missing data
- clustering techniques
- data classification

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 different languages if you prefer. Note, however, that you might be interviewed if we do not understand your code, or if we believe that your code is not running correctly.

2 Data-Sets

2.1 Pulsar data-set

We will use the data-set "(HTRU2)". It is a data-set which describes a sample of pulsar candidates collected during the High Time Resolution Universe Survey [2]. Pulsars are a rare type of Neutron star that produce radio emission detectable here on Earth. They are of considerable scientific interest as probes of space-time, the inter-stellar medium, and states of matter.

Each pulsar produces a slightly different emission pattern, which varies slightly with each rotation. Thus a potential signal detection known as a "candidate", is averaged over many rotations of the pulsar, as determined by the length of an observation. The data set contains $16,259$ spurious examples caused by RFI/noise, and $1,639$ real pulsar examples. Each candidate is described by 8 continuous variables, and a single class variable as detailed below:

- Mean of the integrated profile.
- Standard deviation of the integrated profile.

- Excess kurtosis of the integrated profile.
- Skewness of the integrated profile.
- Mean of the DM-SNR curve.
- Standard deviation of the DM-SNR curve.
- Excess kurtosis of the DM-SNR curve.
- Skewness of the DM-SNR curve.
- Class

HTRU 2 Summary:

- $17,898$ total examples.
- $1,639$ positive examples.
- $16,259$ negative examples.

The data-set can be obtained at `https://archive.ics.uci.edu/ml/datasets/HTRU2`.

2.2 Abalone and mushroom data-sets

For classification tasks we are going to use two data-sets: The "abalone" data-set, where the age of abalone must be estimated from physical measurements; and the "mushroom" data-set, where we will classify whether a mushroom is edible or poisonous (Figure 1). These data-sets are both originally from the UCI Machine Learning Repository [1]. They will allow you to explore classification and other data pre-processing techniques. In particular, the *abalone* data-set requires techniques for handling imbalanced data, while the *mushroom* data-set requires handling missing data.

The *mushroom* data-set, including a detailed description, can be found at `http://archive.ics.uci.edu/ml/datasets/Mushroom`. For the *abalone* data-set, we will use a version especially designed for studying class imbalance, which must be obtained from `http://sci2s.ugr.es/keel/dataset.php?cod=115`. **Note that it is a different version than the original one in the UCI repository.**

Figure 1: Illustration of the abalone and mushroom data-sets.

3 Submission deadline

The deadline for submission is: 4pm, 13 December 2019, Friday. The cut-off deadline is 4pm, 16 December 2019, Monday (with late submission penalty incurred which is one grade or 10%). Submissions after this deadline are not acceptable according to the University regulations.

3.1 Demonstrations:

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.

4 Marking Scheme

The marks will be allocated as follows:

- Overall conclusion and analysis (14%)
- Language (10%)
- Structure and presentation (10%)

Within each of the three parts (Pre-processing, Clustering and Classification) 22% will be allocated based on:

- Working and well annotated code, 8%
- Use of a good variety of methods, 4%
- Results and research, 10%

At the end of this document there is an Appendix explaining what a mark means in Lancaster University and providing 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 can include at most 2 pages of appendix after the main report (in a total of 8 pages).

5 Tasks

5.1 Pre-processing

Pre-processing requires feature selection and/or extraction, and standardising and/or normalising the data. Pre-processing will provide an insight into the structure and dependency of the data, exploring relationships or correlations.

One feature extraction technique is the Principle Component Analysis (PCA). You can extract new, orthogonal (independent) features which are linear combination of the original ones, comment on the amount of variance, interpretability and link with the original features, plot the results using, for example, the one or two of the principle components which contain most of the variance.

Additionally, pre-processing techniques can be applied for handling imbalanced data, and missing values. You will be expected to apply those when handling the Classification tasks.

5.2 Clustering algorithms

Choose at least two clustering algorithms that were studied in the lectures and labs and apply them to the data. 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.

5.3 Classification

In this task you must study the performance of classification algorithms in two challenging datasets: "abalone" and "mushroom". You will need to apply techniques for handling imbalanced data and missing values, and study the impact on the final classification performance. Please study several classifiers, and several techniques for handling class imbalance and missing values.

You must use cross-validation, and present an extensive analysis, exploring different parameters and commenting extensively to demonstrate understanding. Please provide a detailed analysis of the results, e.g., using precision/recall, ROC, classification rate as a function of the size of the training data sub-set, etc.

Compare different classifiers, and the different techniques for handling imbalanced classes and missing data. Please include computational complexity in your analysis (e.g., how much time/memory it takes to train/run a classifier).

6 Additional Comments

You must report in an "acknowledgements" section the use of any libraries, readily available on-line code, and code from on-line tutorials. Additionally, you are free to discuss your work with colleagues, but you must also report in the "acknowledgments" section if anyone is helping 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.

References

[1] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017.

[2] M. Keith, A. Jameson, W. Van Straten, M. Bailes, S. Johnston, M. Kramer, A. Possenti, S. Bates, N. Bhat, M. Burgay, et al. The high time resolution universe pulsar survey–i. system configuration and initial discoveries. *Monthly Notices of the Royal Astronomical Society*, 409(2):619–627, 2010.

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**. Plagiarism can occur on any written work or oral presentation related to your evaluation for a grade. Likewise, using outside help during a test or quiz or other evaluation without the consent of the teacher is an act of cheating.

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.