School of Computing & Communications | Lancaster University

SCC403 – Data Mining

Coursework Assignment

1 Introduction

A data scientist must be able to process various data sets and streams, to know various methods and techniques and be able to select and apply the most suitable algorithms for data mining. In particular, techniques such as data pre-processing, data clustering, and classification.

The objective of the assignment is to conduct data analysis using two sets of different real life data. The first data set concerns the climate and the second data is a video stream. The assignment includes selection and justification of the specific methods for data pre-processing (normalisation, standardisation, feature selection and/or extraction, anomaly detection, missing data (if any)), their implementation and analysis of the results as well as a well annotated code. 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. To achieve top marks a well justified variety of specific techniques is expected. Analysis and understanding of the methods, algorithms and the overall process are the most important elements in addition to the implementation skills such as the code and the presentation.

We expect the use of Python - the most widely used language for machine learning which we also use in the labs, but if you prefer to use a different language we may need to contact you for clarification, if we we believe that your code is not running correctly.

2 Data Pre-processing

2.1 Data Set 1

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

2.2 Data Stream 2

The second data concern a real multi-dimensional video stream showing two moving objects (a car and a motorbike) represented by the file $'OriginalVideoStream.m4v'$. A snap shot of this video stream (a single image frame) is given in Figure 1) which shows a police car in pursuit of a motorcycle. The video contains several multi-channel data sources like RGB (Red-Green-Blue) encoding for each pixel of each frame as well as sound.

Figure 1: An image frame from the original video.

The original video file can be processed using the so-called *background subtraction* method for image processing which results in a binary video (the file $'BinaryVideo.avi'$) where the pixels of the background are black and the pixels of the foreground (moving objects that differ from the background) are white. A snapshot of this video (a binary image frame) is shown in Figure 2.

Within this binary video, of special interest are the foreground pixels and the object that they represent when considered together.

Remember, that *feature extraction* is the process of transformation of the original features (such as pixel colour, e.g. R, G, B or temperatures, pressures, age, etc.) into a set of new, derivative features (e.g. size, shape, area, etc. or principle components). One possible approach for *feature extraction* applicable to Data Stream2 is to form rectangular enclosures (bounding boxes) that surround the suspected objects represented by groups of foreground pixels, see Figure 3.

For example, these can be determined using the top left and bottom right corners of the enclosures (bounding boxes), see Figure 3. Then, based on the coordinates of these corners it is easy to determine the width (W), the length (L), and the area (A) of the rectangles that enclose these objects. For the example provided in Figure 3 it can be derived that:

Figure 2: An image frame with binary info: black - background; white - foreground.

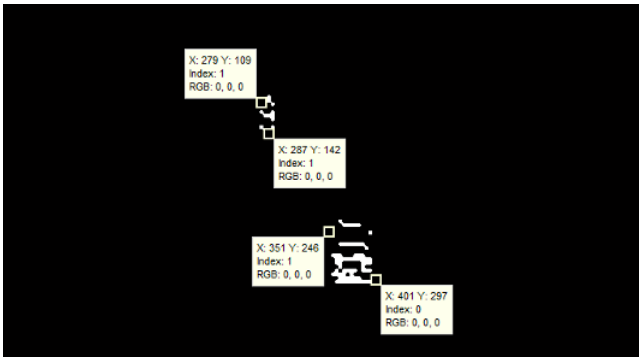

Figure 3: Selecting corner pixels.

- Wmotorcycle = 287 - 279 = 8 pixels
- Wcar = 401 - 351 = 50 pixels
- Lmotorcycle = 142 - 109 = 33 pixels
- Lcar = 297 – 246 = 51 pixels
- A=W*L
- Amotorcycle = 264 pixels^2
- Acar = 2550 pixels^2

Each one of the image frames of the binary video file ($'BinaryVideo.avi'$) where processed as described above and, as a result, the set of W (width), L (length) and A (area) were determined and saved in the file called $'WLA.csv'$. The file has 188 lines and 3 columns. The columns represent the dimensions of the rectangular blob of foreground pixels (W, L and A). Finally, using human expertise (manual annotation) the true labels are provided in the file $'Labels.csv'$ where 1 denotes "car" and 2 - "motorbike". You may notice that the first 16 lines represent image frames in which only the motorbike is visible, while the remaining 172 lines (which represent the next 86 image frames) have both, the police car and the motorbike. So, in total there are 102 image frames in the video-clips.

You should select features based on which the further data processing such as clustering or classification can be performed.

Hint: due to high correlation you may decide to use less than 3 features. Please, justify your choice.

Furthermore, you are expected to apply other pre-possessing techniques and to justify your choice(s).

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.

3 Clustering

The objective is to cluster the climate data set. Choose at least two clustering algorithms and apply them to the climate data set. 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 Classification

The objective is to classify different objects detected from the video stream. Train at least two classifiers of your choice on a part of the data (you may choose what proportion of the data to use for training and what proportion for testing/validation), perform cross-validation and evaluate the performance of the classifiers and report this.

Hints:

1. The first 16 lines of the files $'WLA.csv'$ and $'Labels.csv'$ contain only one of the two objects of interest - only the motorbike - and, therefore, perhaps these 16 cases will not be very useful for training.

2. The remaining 172 lines of the files $'WLA.csv'$ and $'Labels.csv'$ has to be considered in pairs (86 pairs)

3. As you probably realised from Task1 (Data pre-processing) it may not be the best option to use all three values from the file $'WLA.csv'$. Why?

When analysing the performance of the classifiers you should use precision/recall, F1 score and classification accuracy. You may also indicate the time required for training the classifier as a measure of computational complexity (note that the time is always conditional on the type of hardware you use - laptop, computer, CPU/GPU, etc.) and is not an absolute measure, but when making comparisons it can be useful.

5 Marking Scheme

The marks are allocated as follows:

- Structure and presentation (6%)
- Language and style (5%)
- Use of literature and references (5%)

Plus the same for each of the following tasks:

- Data Pre-processing (total 28%)
- Clustering (total 28%), and
- Classification (total 28%)

formed from:

- Level of understanding (6%)
- Depth of analysis (6%)
- Working, well annotated code and results (6%)
- Justification of selected methods (4%)
- 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 (4 pages maximum) following the main report.

6 Deadlines and general requirements

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).

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: 6pm, 16 December 2022, Friday. The cut-off deadline is 6pm, 19 December 2022, 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".

7 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

Student number
line 1: dept. name of organization
line 2-name of the programme and module

***Abstract*— Briefly describe the outline of your report. You can download a template (Word or LaTex from** https://www.ieee.org/conferences/publishing/templates.html **)**

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.