Data Mining COMP3340/ COMP6340

Assessment 1 - Part 1 of 3

Deadline: 3rd September (Sunday) 23:59PM via Canvas

1. General Guidelines

Assessment 1 includes writing and programming components. This assessment will be divided into **3 Parts**, and the objective is to develop skills in the area of Data Mining and Data Analytics. A signed cover letter **must** be included with the submission, so it can be marked. Submission of each part will be made via Canvas.

Assessment 1 - Part 1 is due on 3rd September (Sunday), 23:59PM and submission must be done via Canvas.

1.1 Marks

The whole Assessment accounts for **50 marks**. Please find below the distribution of marks for each part of the 3 parts:

- **Part 1: 10 marks**
- Part 2: 15 marks
- Part 3: 25 marks

1.2 Goals

This assessment, in its three parts, aims at providing students with some hands-on experience in the analysis of real-world datasets. Some of the datasets will be discussed in class, others will be selected by the students to work on them and report on the results.
The purpose of the *Assessment 1- Part 1* it to start the set of activities and commence working during the semester leading to the Programming Project and Report. It is meant to be guiding you for the development of a term project in data mining.

1.3 Datasets

It is reasonable then to start by referring to several datasets that are available and you can use during the course. Some datasets are too small but they are good to start testing ideas/code, the performance of algorithms, etc. Please find below, on Table 1, some examples of datasets that were related to the Lecturer's previous research and that will be discussed in class.

Table 1: Dataset examples.

Classification Datasets	Regression Datasets
Fisher's Iris Flower[1]	Concrete Compressive Strength Dataset
US Presidency[2]	Airfoil Self-Noise
Alzheimer's Disease[3]	Boston Housing
Customer Churn[4]	Yacht Hydrodynamics
Tripadvisor[5]	Energy efficiency of Buildings *(two datasets, one for 'Heating' and the other for 'Cooling' loads)*
Amazon Co-purchasing and Product[6]	
Students' Academic Performance[7]	

[1]Fisher's Iris Flower dataset is very famous and references to it possibly appear in many textbooks about data mining and machine learning. This dataset was introduced by Ronald Fisher in 1936 (see: https://en.wikipedia.org/wiki/Iris_flower_data_set).

[2]This dataset was first contributed by Lichtman and Keilis-Borok in 1981 and it is available on the public domain. See Table 2 from:

> *Pablo Moscato, Luke Mathieson, Alexandre Mendes, and Regina Berretta. 2005. The electronic primaries: predicting the U.S. presidency using feature selection with safe data reduction. In Proceedings of the Twenty-eighth Australasian conference on Computer Science - Volume 38 (ACSC '05). Australian Computer Society, Inc., AUS, 371–379.*
> *https://dl.acm.org/doi/10.5555/1082161.1082202*

We have presented the *k-Feature Set* problem using this dataset in that publication. You will use it to find association rules, for instance, to create proximity graphs, to test classification algorithms, etc.
The dataset is also discussed in: https://www.springerprofessional.de/en/marketing-meets-data-science-bridging-the-gap/16761198. Remember that you have access to this chapter via the Library (the entire book is available online).

[3]This dataset was first contributed by Ray et al., in 2007 in:

I. *Classification and prediction of clinical Alzheimer's diagnosis based on plasma signaling proteins*, Ray et al., Nature Medicine 13, 1359 - 1362 (2007). Published online: 14 October 2007 | doi:10.1038/nm1653 http://www.nature.com/nm/journal/v13/n11/abs/nm1653.html

If you go to the URL above, the dataset is available via the Supplementary Material. Following the link `Supplementary Info' you will get more information about methods and the dataset as an Excel file. The Excel file contains a sheet with a `Training Set', a `Test Set' in another. There is also a sheet containing samples belonging to people who have presented Mild Cognitive Impairment (MCI) that then "converted" to Alzheimer's Disease (while others converted to Other Dementias (OD)).
We have used this dataset in other two research papers that I also recommend you to review:

I. Identification of a 5-Protein Biomarker Molecular Signature for Predicting Alzheimer's Disease Gómez Ravetti M, Moscato P (2008). PLoS ONE 3(9): e3111. doi:10.1371/journal.pone.0003111 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0003111
II. Differences in Abundances of Cell-Signalling Proteins in Blood Reveal Novel Biomarkers for Early Detection Of Clinical Alzheimer's Disease. Rocha de Paula M, Gómez Ravetti M, Berretta R, Moscato P (2011). PLoS ONE 6(3): e17481. doi:10.1371/journal.pone.0017481 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017481

[4]A customer who is transferring its demands from one service provider to another is said to be "churning". The "churn rate" is of a concern to businesses. Both marketing specialists and business operators would like to know the reasons why this is happening and identify what can be done to reduce the rates. We suggest you also consider the following paper and its associated dataset.

I. Customer churn prediction using a hybrid genetic programming approach. Obiedat, R., Alkasassbeh, M., Faris, H., Harfoushi, O.: Scientific Research and Essays 8, 1289–1295 (2013) http://www.academicjournals.org/journal/SRE/article-full-text-pdf/D65F3CB28782
II. A Genetic Programming Based Framework for Churn Prediction in Telecommunication Industry. Faris H., Al-Shboul B., Ghatasheh N. (2014). In: Hwang D., Jung J.J., Nguyen NT. (eds) Computational Collective Intelligence. Technologies and Applications. ICCCI 2014. Lecture Notes in Computer Science, vol 8733. Springer, Cham http://link.springer.com/chapter/10.1007/978-3-319-11289-3_36

[5]Customers that purchase more than two items at the same time may give some interesting insights for brand developers. We have studied such a behaviour in a recent publication.

I. Where Does My Brand End? An Overlapping Community Approach. Gabardo A.C., Berretta R., de Vries N.J., Moscato P. (2017). In: Leu G., Singh H., Elsayed S. (eds) Intelligent and Evolutionary Systems. Proceedings in Adaptation, Learning and Optimization, vol 8. pp 133-148 Springer, Cham

http://link.springer.com/chapter/10.1007/978-3-319-49049-6_10

If you are interested in accessing co-purchasing data available online please check: Amazon Product Data http://jmcauley.ucsd.edu/data/amazon/links.html

[6]The Tripadvisor dataset was first published by Honging et al. and it includes over 235,793 hotel reviews collected from Tripadvisor in 2009. It includes written text reviews as well as scaled responses on certain aspects. The aspects on which Tripadvisor users rated the hotels on a 5-point scale are as follows; *'Value', 'Room', 'Location', 'Cleanliness', 'Check-in/Front Desk', 'Service'* and *'Business Service'*. Tripadvisor and tourism product and service providers may find useful business insights from analysing large amounts of customer review data.

I. Latent Aspect Rating Analysis without Aspect Keyword Supervision. Hongning Wang, Yue Lu and ChengXiang Zhai. The 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2011), P618-626, 2011.
II. Latent Aspect Rating Analysis on Review Text Data: Hongning Wang, Yue Lu and Chengxiang Zhai. A Rating Regression Approach. The 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD'2010), p783-792, 2010. http://times.cs.uiuc.edu/~wang296/Data/

[7]The Students' Academic Performance Dataset was the subject of a Kaggle competition and can be downloaded from here: https://www.kaggle.com/aljarah/xAPI-Edu-Data. This dataset has been studied in at least two publications:

I. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2016). Mining Educational Data to Predict Student's academic Performance using Ensemble Methods. International Journal of Database Theory and Application, 9(8), 119-136.
II. Amrieh, E. A., Hamtini, T., & Aljarah, I. (2015, November). Preprocessing and analyzing educational data set using X-API for improving student's performance. In Applied Electrical Engineering and Computing Technologies (AEECT), 2015 IEEE Jordan Conference on (pp. 1-5). IEEE.

1.3.1 Other Datasets of Interest

- Datasets for Data Mining and Data Science – Kdnuggets www.kdnuggets.com/datasets/index.html
- Adaptive Predictive Modelling *(an interesting book with datasets useful for classification but also for regression problems)* http://appliedpredictivemodeling.com/data/
- Datasets for Data Mining *(from The University of Edinburgh, Scotland)* http://www.inf.ed.ac.uk/teaching/courses/dme/html/datasets0405.html

- IBM's Guide to Sample Datasets*
  https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/
  **Note: See the discussion in the Forum, Can you use these datasets?*
  https://community.watsonanalytics.com/discussions/questions/22783/free-to-use-the-sample-dataset-anyway-i-want.html
- Stanford Large Network Dataset Collection: http://snap.stanford.edu/data/
- Free Datasets (Courtesy of RDataMing.com):
  http://www.rdatamining.com/resources/data

1.3.2 Additional Topics of Interest

Please discuss with the Course Coordinator during the week starting the possible topics you have an interest in (either in class or via email) to select another dataset for you to work in future Assessments.

1.4 Ethical use of Information

When selecting a dataset for use, you will need to check if it can be used for academic purposes. A common misconception exists and students and staff at academic universities should be aware of it. The argument is a bit like this:

~~*"If a dataset is de-identified and it is available online, then I can use it for research purposes."*~~

In fact, **that is not the case** at The University of Newcastle. The University of Newcastle Human Research Ethics Committee (HREC) operates in accordance with the NHMRC National Statement on Ethical Conduct in Human Research (NHMRC's): https://www.nhmrc.gov.au/guidelines-publications/e72

It is recommended that you to read the University webpage on *"What needs ethics approval?"* http://www.newcastle.edu.au/research-and-innovation/resources/human-ethics/what-needs-ethics-approval. See in particular *point 1.4* in the link just highlighted:

"Ethics approval must be sought for research involving human participants.
A 'participant' is someone who:
…
4. Is identified or de-identified in data banks or unpublished human research data, e.g. an analysis of existing unpublished data collected by another researcher or collected for a different research project."

Please note that it says that *"Ethics approval* ***must be sought****…".* Contact your Course Coordinator or the Research Ethics Advisor of your Faculty if you have any questions.
https://www.newcastle.edu.au/research-and-innovation/resources/human-ethics/research-ethics-advisors-reas

2. Assessment 1 – Part 1

2.1 Marking Criteria

Each successful or in-depth attempt to complete each one of the exercises will attract one or two points (depending on the task). All results would need to be uploaded as a single *.pdf* file in Canvas. Include a copy of the script/software, screenshot etc. used to accomplish the task.

2.2 Tasks

Exercise 1 (1 mark) In João Moreira's book, *Hamming Distance*, a metric for comparing two binary data strings is discussed. Using the *US Presidency Dataset* presented previously*:*

a) Write a computer program that computes the *Hamming Distance* matrix between all pairs of elections and the matrix generated by the *Jaccard measure*.

b) Same as (a) above, but in this case you will compute the matrices that are obtained by comparing columns (i.e. attributes) of the *US Presidency Dataset.*

c) Now, Calculate the *Minimum Spanning Tree* (having as input that distance matrix). Visualize the result using the yEd software.

Exercise 2 (1 mark) Same as Exercise 1 (c), but now calculate the *Relative Neighborhood Graph* instead of the *MST*.

Exercise 3 (1 mark) Same as Exercise 1 (c), but to calculate the MST given as input a distance matrix generated by the *Jaccard measure*.

Exercise 4 (1 mark) Same as Exercise 3, but now calculate the RNG instead.

Exercise 5 (1 mark) Using the *Hamming Distance* matrix between all pairs of elections of the *US Presidency Dataset*, calculate the *k-NN* graph with *k=2*. Visualize the result using the yEd software.

Exercise 6 (1 mark) Using the *Jaccard measure* to generate a distance matrix between all pairs of elections of *US Presidency Dataset*, calculate the *k-NN* graph with *k=2*. Visualize the result using the yEd software.

Exercise 7 (2 marks) Based on the results of Exercise 5 and Exercise 1, identify the edges of the MST that are also part of the *k-NN* graph. Visualize the resulting graph using the yEd software. Write down the elections in tabular form grouped by the different clusters observed.

Exercise 8 (2 marks) Based on the results of Exercise 6 and Exercise 3, identify the edges of the MST that are also part of the *k-NN* graph. Visualize the resulting graph using the yEd software. Write down the elections in tabular form grouped by the different clusters observed.

Note: The Jaccard measure is a measure of similarity, you may need to define your measure of distance ("dissimilarity") as a function of it.