Data Mining COMP3340/ COMP6340

Assessment 1 - Part 3 of 3

Deadline: November 11th 11:59PM via Blackboard

1. General Guidelines

As mentioned in previously, *Assessment 1* includes writing and programming components. This assessment is divided into **3 parts**, and you should have completed **Part 1** and **Part 2** by now. The main objective of this assessment 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 (for each part of the assignment).
Submission of each part will be made via Blackboard.

Assessment 1 - Part 3 is due on **November 11th, 11:59PM**, and submission must be done via Blackboard as a single PDF file[1].

[1] Please include all scripts, graphs, figures, etc. in the PDF (for instance as annexes).

1.1 Marks

The whole *Assessment 1* accounts for **50 marks**. Please find below the distribution of marks for each 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 3* is to continue the set of activities and work you have been doing during this semester, leading to the Programming Project and Report. It is meant to continue (following *Assessment 1- Part 1* and *Part 2*) to guide you for the development of a term project in data mining.

1.3 Datasets

For *Assessment 1 - Part 3* you may need to use some of the datasets used previously for *Part 1* and *Part 2*. Please refer to the Dataset Section of previous assessment sheets if necessary. You are also required to do some analysis on a dataset of your interest. Please refer to *Part 1* of the assessment for insights on datasets, if you do not have any in mind.

1.4 Programming Language and Visualization Tools

For the purpose of completing this assignment, any programming language may be used (we recommend *Python* or *R*).

Graphs can be nicely displayed with a variety of different software packages. You can choose anyone you want. We also recommend to have a look at the functionalities provided by *yED – Java Graph Editor*. The package is available from:
http://www.yworks.com/en/products_yed_about.htm

Another useful package is *Graphviz - Graphs Visualisation Software*, available from:
http://www.graphviz.org/

2. Assessment 1 – Part 3

Assessment 1 – Part 3 is composed of a total of 12 tasks. Remember, as you did previously for *Part 1* and *Part 2* of the assessment, all results for *Part 3* must be uploaded as a single *.pdf* file on Blackboard. Include a copy of the script/software, screenshot etc. used to accomplish the task in the same PDF file.

2.1 Marking Criteria

Each successful or in-depth attempt to complete each one of the exercises will attract **one (1), two (2), three (3), or (4) points**, depending on the task. Mark allocation and notes about the marking criteria is specified in each task.

2.2 Tasks

Exercise 1 (2 marks). In *Assessment 1- Part 1* we have introduced a dataset for the classification and prediction of clinical Alzheimer's Disease first reported by Ray et al. in 2007. In *Assessment 1 - Part 2* your task was to generate a *Gabriel Graph (GG)* for samples (and also for proteins) and the *Relative Neighbourhood Graph (RNG)* for samples and also for proteins.

Using the same algorithm/program and dissimilarity metric you have developed before, find the *Relative Neighbourhood Graph (RNG)* for a regression dataset[2] of your choice. Visualise the result using a method of your choice. It should have several features/variables and a reasonably large number of samples (use your judgement, neither too small nor too large that you can't generate the graph).

Note: You may or may not need to discretise the values of the features. If you need to do that, we would suggest to use the Fayyad-Irani discretisation method. A tutorial about this method is available under Assessment -> Assessment - Part 2 -> Fayyad-Irani's Discretisation folder. If you use another one, please explain it and give references of its previous use in the literature. You can also use any of the datasets from Section 1.3: Datasets.

[2] The Concrete Compressive Strength dataset of the UCI Machine Learning repository in https://archive.ics.uci.edu/ml/datasets.php is a good example of such a dataset. However, the number of samples is relatively big (about a thousand) and the computation of the *RNG* may take some time. However, computational efficiency is not an issue for obtaining marks here, but it would be good also to report the time your algorithm has taken to obtain the result. You can also use any of the datasets discussed in *Assessment - Part 1*.

Exercise 2 (3 marks). In *Assessment 1 - Part 2,* you have implemented a feature selection technique that, using the Alzheimer's Disease Training set available, selected a subset of features (from the total of 120 measured proteins). Your task was to implement a classifier system that "learned" from the Training set and then you calculated the performance of your classifier according to Sensitivity, Specificity, Accuracy, F1-score, Matthews Correlation Coefficient and Youden's J statistic measures.

You now need to use that *feature selection technique* on a classification dataset of your choice. The classification problem could be of a binary type (as it was the one of the Alzheimer's Disease problem[3]). First, you need to clearly identify which samples correspond to your Training set and which are part of the Test set.

Note: You will obtain one mark for the successful computation of the task, and two additional marks if you clearly explain all the procedures in your own words and it is at a standard that can be reproduced from the text.

[3] You may or may not need to discretise the values of the features. If you need to do that, we would suggest to use the Fayyad-Irani discretisation method. If you use another one, please explain it and give references of its previous use in the literature. You can also use any of the datasets discussed in Assessment 1 - Part 1.

Exercise 3 (2 marks). "<u>You are the teacher now!</u>": Write down the formal mathematical definition of the *k-Feature Set problem*, and then show with a very small example (designed by you) of a problem instance that has a *3-Feature Set* as a solution, but no *2-Feature Set*. Also, the *3-Feature Set* solution is a feature set that separates the set of samples, but it does not admit linear separability.

Note: You will obtain one mark for correctly presenting the mathematical definition of the problem (no mistakes!) and another mark for an example that satisfies the requirements of the exercise.

Exercise 4 (3 marks). "<u>You are the teacher again!</u>": Write down the formal mathematical definition of the *l-Pattern Identification Problem* and then show with a very small example (that you design) of a problem instance that has a *4 patterns* that do the job (meaning that there is a solution for the problem), but no *3-patterns*.

Note: You will obtain one mark for correctly presenting the mathematical definition of the problem (no mistakes!) and another two marks for an example that satisfies the requirements of the exercise.

Exercise 5 (2 marks). Based on *Exercise 1* and *Exercise 2* of *Assessment 1 – Part 1*, given a regression or classification dataset of your choice, and a dissimilarity measure of your choice, follow the three steps to generate a clustering result:

a) Calculate the *Minimum Spanning Tree (MST) graph* of the samples;
b) Calculate the *k-NN graph*;
c) Identify the edges of the *MST* that are also part of the *k-NN graph*;
d) Report the results either via a graph format (e.g. *yEd* of *Gephi*) or using a table where each row corresponds to a cluster (may include cluster ID) and rows correspond to the samples which are the member of that cluster.

Assuming that you have found an appropriate value for k (of your choice) you will have now your samples separated in different clusters, each of the trees obtained after following the steps (a), (b), and (c) above.

Note: You will obtain one mark for correctly executing steps (a), (b) and (c), and another mark for presentation of the clustering result (d), depending on the quality of representation shown.

Exercise 6 (3 marks). Using the same dataset you have selected for the previous *Exercise 5*, and an unsupervised clustering algorithm of your choice, separate the samples in more than 2 different clusters. The purpose of this exercise is to compare the results you have obtained in *Exercise 5* (with the *MST-kNN approach*) with the ones you have obtained with your alternative method of choice. After doing that, proceed the following sub-tasks:

a) Define what an *inter-rater reliability method* is and give examples of several types of them;
b) Identify which *inter-rater reliability method* can be used to compare the results of the two types of clustering;

c) Use one *inter-rater reliability method* of your choice to compare the results of both clusterings[4].

Note: You will get a mark for each of the steps (a), (b), and (c) being correctly executed. In case you are not able to complete Exercise 5, you may use another clustering algorithm that is not the MST-kNN, but a mark will be deducted from the total.

[4] Note: You can use any method, but due to (b) above you need to clearly justify your selection. This Exercise does not restrict you to the type of inter-rater reliability method.

Exercise 7 (2 marks). About classification, answer the following questions:

a) What the concept of *"lazy classification"* means in data mining and what are the problems associated with *"class imbalance"*;
b) Show, with at least three examples of several confusion matrices, how we can obtain relatively good results for *sensitivity*, *specificity* and *accuracy*, and also compute the *Matthews Correlation Coefficient* behaviour in those cases.

Note: One mark for each of the items (a, and b) correctly completed and for good explanations of the behaviours obtained.

Exercise 8 (2 marks). Given a set of points in the *Euclidean* plane in two dimensions (corresponding to a set of objects for which two measures are given), you can cluster these objects using several algorithms/heuristics. For instance, you can use the k-means algorithm, hierarchical clustering algorithms, and the previously discussed *MST-kNN algorithm*. Your task is, using a dataset of your choice, present at least three clustering algorithms that produce significantly different results. You need to illustrate the findings of the chosen algorithms.

Note: You will get one mark for doing the exercise correctly (algorithm, results, etc.), and an extra mark for presentation of findings and clarity of explanation.

Exercise 9 (3 marks). Consider the US Presidency dataset (you are already familiar with it throughout the lecture, workshops and previous assessment tasks) for market basket analysis using Apriori Algorithm. You have to consider each year as a transaction of the market basket. Each of the questions in the dataset should be considered as Item and the answer of the questions *(Q1 to Q12)* of *0* as *"absence"* and *1* as *"presence"* of the item in the transaction. In

.arff format required for weka, "?" is used to represent the absence/missing item in the transaction description. For example, the transaction id 1864 will be treated as *{Q5, Q8, Q9}* as per the following sample.

Here, you have to separate the dataset into two groups, one for each of the *Target* value *(1=incumbent, 0=challenger)* of the victory, let assume the name as *US-Incumbent.arff* and *US-Challenger.arff*. You have to process the dataset suitable for Market Basket analysis using either Weka Software or the from Python (python-weka-wrapper package) or R (RWeka library) and answer the following questions:

a) Run *Apriori* for Class Association Rule (by setting the parameter *car=true*) for the *US-Incumbent.arff* and *US-Challenger.arff* Victory. What is the maximum size of the *itemset* found by Weka for both cases? Report the largest *Frequent itemset* found by Weka for each set. Compute the *confidence* value and *lift* score of the top 3 association rules found by the *Apriori* algorithms for each of the datasets.

b) Clearly explain the step-by-step process of finding association rules using the *Apriori* algorithm on the US presidency dataset for *Incumbent victory* and *Challenger Victory* using the *support threshold* to *60%*.

Note: You will get one mark if you correctly answer the question (a), and two marks for correctly answer question (b). In question (b) you will get one mark for showing the steps of Apriori correct and another for clarity of explanation.

Exercise 10 (1 mark). For the US presidential election dataset that was amply discussed in class, workshops and previous assessments, and your current knowledge about it (association rules, feature sets, etc.), is it possible to use feature engineering to come up with a set of new features such that the problem becomes linearly separable?

Note: You get a point if you correctly answer the question, meaning that you provide a linear separation (by engineering a new set of features) or, alternatively, you prove that it is not possible to do so.

Exercise 11 (1 mark). Again, in regards to the same dataset from the previous exercise, give an example of a *decision tree induction algorithm* that classifies all the elections.

Note: You get a point if you correctly answer the question. Show how the tree correctly classifies all samples.

Exercise 12 (1 mark). Define the following concepts:

a) Cross-validation

b) Bootstrapping

c) Imputation

Note: You get the mark if you correctly define all three concepts without mistakes (0.33 each).