COMP3420 – Advanced Databases and Data Mining
Assignment 2 Due Thursday 19 May 2016, 5 pm

Last update 1 May 2016

Please notify Peter Christen (email `comp3420@anu.edu.au`) if anything is unclear to you or if you find mistakes.

Objectives

The objectives of this second assignment are to apply the topics learned in the middle part of the course by (1) **manually** conducting association rules mining on a small database; (2) describing certain characteristics of different clustering algorithms; (3) calculating classifier accuracy measures; (4) answering questions on decision trees; and (5) carrying out a small data mining project in **Rattle**.

This assignment is worth 20% of your total course mark. It will be marked out of 20 as indicated below. The estimated time we expect you to spend on this assignment is roughly 20 hours in total. (1 hour per mark).

Submission

Submission will be done using Wattle. Click on the link Assignment 2 submission (to be made available) in topic in week 12 (16 to 22nd May) to upload your file.

You may upload as many draft submissions as you want. However, make sure to submit a final version before the deadline (5pm Thursday 19 May 2016). If you do not do this your submission will remain in draft and may not be marked!

You will have to **submit one PDF file that contains five parts**:

- The answers to the question on association rules mining given in task 1 below.
- The answers to the question on characteristics of clustering algorithms given in task 2 below.
- The answers to the question on classifier accuracy measures given in task 3 below.
- The answers to the question on decision tree classification given in task 4 below.
- A report that contains the details of your data mining project as described in task 5 below.

Important

- Make sure that your submitted PDF file contains **on the first page your ANU ID only, NOT your name!**
- **We only accept PDF documents**, no other formats such as Word or OpenOffice documents.
- We do not accept handwritten submissions (in any part).
- **Name your submitted file: u1234567-ass2.pdf** (replace '1234567' with your ANU ID).
- The **maximum total length of your submission must not be more than ten (10) A4 pages.**
- You have to use a **font size of at least 12 points.**
- The file must be **submitted by 5 pm on Thursday 19 May 2016.**

Extensions

Students will only be granted an extension on the submission deadline in exceptional circumstances. Work and sporting commitments are normally NOT sufficient grounds. If you think you have grounds for an extension, you should notify the course coordinator (email `comp3420@anu.edu.au`) as soon as possible and provide written evidence in support of your case (for example a medical certificate). Peter will then decide whether to grant an extension and inform you as soon as practical.

Penalties

Following the new ANU wide late penalty policy, **NO late submissions can and will be accepted after the deadline.** This means the submission system will close at 5 pm on Thursday 19 May 2016.

Penalties for submissions that are longer than ten (10) pages are as follows:

Number of pages	11	12	13	14 or more
Penalty from 20 marks	-2	-4	-6	-8

If you use a font size smaller than 12 points we will penalise your submission by deducting 2 marks.

Plagiarism

No group work is permitted for the assignment. We do encourage you to discuss your work in the labs and lectures, but we expect you to do the assignment work by yourself.

Make sure you read the COMP3420 course administrative handout, especially the section on Plagiarism:

`https://docs.google.com/document/d/1-L0ssIVJmEsWNMNUS8zXKe9z9It_ccLDfmBQWDdY2vo/edit?pref=2&pli=1`

Make sure to also look at the information given on the ANU Research School of Computer Science Current Students page:

`https://cs.anu.edu.au/cs-current-students/cs-undergraduate`

The ANU Web site on Academic Honesty and Plagiarism also contains excellent resources on this topic:

`http://academichonesty.anu.edu.au/`

If you do include material from some other documents (e.g. graphics and figures, tables or formulas extracted from a paper, a book or a Web site), then you clearly have to make attribution, for example by writing the name of the paper, book, Web site, etc. where you got it from, or by adding a reference to the material/source into your report.

Tasks

1. **Association rules mining** (5 marks)

   In this task, you have to **manually** create a small database with ten transactions as specified below, and then **manually** find all large (frequent) item-sets of length 2, 3 and 4, and rules of length 3 for the **first two alphabetically sorted** large item-sets of length 3, and calculate their support, confidence and lift values.

   The minimum support to be used is 3.

   Generate the ten-transaction database as follows:

   - The first seven transactions are given here:

T1:	a, b, e, f
T2:	a, b, c, e, f
T3:	c, d, e
T4:	b, c, e
T5:	a, b, c, e, f
T6:	b, c, e
T7:	b, c, f

   - The last three transactions you need to generate **based on your ANU ID** as described below using the following table.

Digit	Third Last	Second Last	Last
0	a, b, f	a, b, d	c, e
1	a, c, d	a, c, f	**b, c, d, f**
2	a, b, c, d	b, c, e	a, c, e
3	a, b, e, f	a, c, d	a, b, c
4	**a, b, d, f**	**c, d, e**	a, d, e
5	c, e, f	a, b, c, f	a, b, c, e
6	a, b, e, f	b, d, e	b, e
7	a, b, c, f	a, c, e	a, d, e
8	b, c, d, f	c, e, f	a, b, c
9	a, b, d, f	a, c, d	a, b, e

For each of the last three digits of your ANU ID, select the corresponding item-set from the above table.

For example, for the ANU ID u2345**441**, the following three item-sets (transactions) would be selected from the table (shown in bold in the table above):

T8:	a, b, d, f	*First column in row 4 for third last digit in 2345441*
T9:	c, d, e	*Second column in row 4 for second last digit in 2345441*
T10:	b, c, d, f	*Third column in row 1 for last digit in 2345441*

Based on the 10-transaction database you have now generated, conduct the following:

(a) Find **all the large (frequent) item-sets of length 2** that have a count of at least 3 transactions (i.e. minimum support 3).

(b) Find **all the large (frequent) item-sets of length 3** that have a count of at least 3 transactions (i.e. minimum support 3).

(c) Find **all the candidate item-sets and the large (frequent) item-sets of length 4** that have a count of at least 3 transactions (i.e. minimum support 3). Describe why pruning of candidate sets occurs (or why not if no candidate sets are pruned).

(d) **Sort the large item-sets of length 3 alphabetically** (according to the items they contain), and **for the first two of them**, generate the rules of length 3 they contain (that have two items on the left and one on the right hand-side). For each of these rules calculate their support and confidence (as percentage numbers between 1% and 100%, rounded to one digit after the decimal point), and their lift (rounded to two digits after the decimal point).

The output you have to write into your assignment report should follow the example given below for ANU ID u2345**441**:

```
Task 5 of assignment 1 (5 marks):
=================================

(1) Database:  Trans ID | Item-set
              ---------------------------------
                    T1  | a, b, e, f
                    T2  | a, b, c, e, f
                    T3  | b, c, e
                    T4  | a, b, c, e, f
                    T5  | b, c, e
                    T6  | c, d, e
                    T7  | b, c, f
                    T8  | a, b, d, f           [digit from ANU ID: 4]
                    T9  | c, d, e              [digit from ANU ID: 4]
                   T10  | b, c, d, f           [digit from ANU ID: 1]


(2) Large 2 item-sets:  Item-set   | Count
    (0.5 mark)          ------------------
                          a, b     |   4
                          a, e     |   3
                          a, f     |   4
                          b, c     |   6
                          b, e     |   5
                          b, f     |   6
                          c, d     |   3
                          c, e     |   6
                          c, f     |   4
                          e, f     |   3
```

```
(3) Large 3 item-sets:  Item-set     | Count
    (0.5 mark)          --------------------
                          a, b, e    |   3
                          a, b, f    |   4
                          a, e, f    |   3
                          b, c, e    |   4
                          b, c, f    |   4
                          b, e, f    |   3


(4) Candidate 4 item-sets:  Item-set
    (0.5 mark)              ---------------------
                            a, b, e, f
                            b, c, e, f  <- Pruned in Apriori prune step

    Explain the pruning of candidate sets, or why no pruning occurs.
    (0.5 mark)

    Large 4 item-sets:  Item-set       | Count
    (0.5 mark)          ----------------------
                         a, b, e, f   |   3

(5) Frequent rules of length 3 from first two large 3 item-sets:
    (1 mark for correct rules and their support, 0.5 mark for
     correct confidence, and 1 mark for correct lift)
            Rule        | Support | Confidence |  Lift
            -------------------------------------------
            (a, b) -> e  |   30.0  |      75.0  |  1.07
            (a, e) -> b  |   30.0  |     100.0  |  1.25
            (b, e) -> a  |   30.0  |      60.0  |  1.50
            (a, b) -> f  |   40.0  |     100.0  |  1.67
            (a, f) -> b  |   40.0  |     100.0  |  1.25
            (b, f) -> a  |   40.0  |      66.7  |  1.67
```

You will receive the following marks as indicated in the above example:

- 0.5 mark for the correct calculation of the large item-sets of length 2.
- 0.5 mark for the correct calculation of the large item-sets of length 3.
- 0.5 mark for the correct calculation of the candidate item-sets of length 4.
- 0.5 mark for an explanation of the pruning or why no pruning occurs.
- 0.5 mark for the correct calculation of the large item-sets of length 4.
- 1 mark for the correct rules generated and their correct support values.
- 0.5 mark for the correct confidence values for these rules.
- 1 mark for the correct lift values for these rules.

If you do not calculate your transaction data set correctly you will be penalised 1 mark.

---

2. **Characteristics of clustering algorithms** (4 marks)

Describe each of the following clustering algorithms in terms of the following criteria: (1) the shapes of clusters that can be determined; (2) the required input parameters; (3) any limitations; (4) an example application where the algorithm would be particularly well suited to the task and why it is appropriate.

(a) AGNES

(b) CLARA

(c) DBSCAN

(d) k-means

Your answer to each technique is worth 1 mark (4 marks in total).

3. **Classifier accuracy measures** (4 marks)

This third task consists of two parts.

(a) In the first part you have to manually calculate several accuracy measures based on a confusion (or error) matrix which is assumed to come from a binary (2-class) supervised classifier.

The confusion matrix is constructed as follows:

(1) Take the seven digits of your ANU ID.
(2) The number of **true positives** (TP) are the first five (5) digits of your ANU ID.
(3) The number of **false positives** (FP) are the last five (5) digits of your ANU ID.
(4) The number of **true negatives** (TN) are the last six (6) digits of your ANU ID.
(5) The number of **false negatives** (FN) are the first three (3) digits of your ANU ID.

Once you have these four numbers (TP, FP, TN, and FN), you have to complete (and write into your assignment report to be submitted) the following tasks:

(1) Fill-in a confusion matrix similar to the example given below. You also need to write down the 'Total' number (sum of the four entries in the confusion matrix).
(2) Calculate a second normalised confusion matrix, where the numbers are proportions of the total number. Round the numbers to five (5) digits after the decimal point.
(3) Calculate the accuracy and error rate as percentages (rounded to two (2) digits after the decimal point) based on the confusion matrix from (1).
(4) Calculate the specificity, precision and recall as percentages (rounded to two (2) digits after the decimal point) based on the confusion matrix from (1).
(5) Calculate the f-measure (also called f-score or F_1 score) of precision and recall as percentages (rounded to two (2) digits after the decimal point). To get the formula for f-measure you will need to do some research. **Make sure to add a reference to the source of where you got the formula for f-measure.**

For points (3) to (5) you must show your workings (formulas used and how you calculated the result). If you only provide the numerical result values you will not get any marks.

Calculate the result values for the measures in points (3) to (5) based on the total values from the confusion matrix from (1).

The output you have to write into your assignment report should follow the example (for ANU ID u1234567) given below:

```
Uni ID: 1234567
  TP = 12345
  FP = 34567
  TN = 234567
  FN = 123

1.   Confusion matrix:
     (0.5 mark)  -----------------------
                 | Pred Pos | Pred Neg |
                 |----------+----------|
        True Pos |   12345  |     123  |
        True Neg |   34567  |  234567  |
                 -----------------------  Total: 281602

2.   Normalised confusion matrix:
     (0.5 mark)  ---------------------------
                 |  Pred Pos   |  Pred Neg  |
                 |------------+------------|
        True Pos |   0.04384   |   0.00044  |
        True Neg |   0.12275   |   0.83297  |
                 ---------------------------

3.  (1 mark, 0.5 each):
   Accuracy =    [show your workings] = 87.68%
   Error rate =  [show your workings] = 12.32%

4 and 5.  (1 mark, 0.25 each):
   Specificity = [show your workings] = 87.16%
   Precision =   [show your workings] = 26.32%
   Recall =      [show your workings] = 99.01%
   F-measure =   [show your workings] = 41.58%
```

You will receive 0.5 marks for a correct confusion and normalised confusion matrix each, 0.5 marks each for correct accuracy and error rate (with correct workings shown), and 0.25 marks each for correct specificity, precision, recall and f-measure (with correct workings shown).

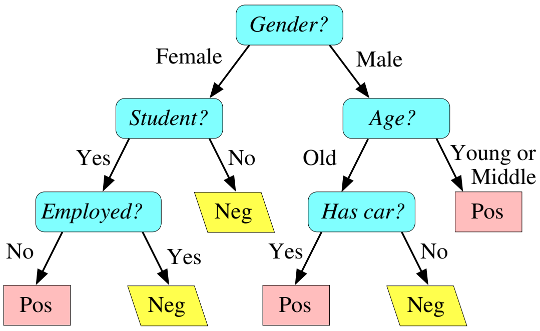
(b) Looking at your numbers of TP, TN, FP and FN, is this a balanced or imbalanced classification problem? Explain why? Calculate and write the balance of the problem (as a ratio) into your assignment submission.

Describe in one or two sentences which of the measures calculated in (a) above is/are suitable for this classification problem or not, and explain why.

You will receive 0.5 mark for correct balance (ratio) value and 0.5 mark for a good explanation of which measure(s) to use or not to use.

4. **Decision tree classification** (2 marks)

Assume the following decision tree has been generated on a set of training records:



For the four following test records, answer if they are a **true positive** (TP), a **false positive** (FP), a **true negative** (TN), or a **false negative** (FN). You must explain your answers. You will not receive any marks if you only write TP, FP, etc.

(a) Age=Young, Gender=Male, Employed=Yes, Student=No, Has car=No, Class=Neg

(b) Age=Old, Gender=Male, Employed=Yes, Student=Yes, Has car=No, Class=Neg

(c) Age=Young, Gender=Female, Employed=Yes, Student=No, Has car=No, Class=Neg

(d) Age=Middle, Gender=Female, Employed=Yes, Student=Yes, Has car=Yes, Class=Pos

You will receive 1/2 mark for each correct answer.

5. **Data mining project** (5 marks)

Conduct a data mining project using **Rattle** on a data set of your choice from the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/datasets.html`). You will have to write and submit a report which details the steps you have done in your data mining project.

You must provide the name of the data set and where you got it from (the URL to the page where it is available).

Note: you are not allowed to use any of the data sets used in the COMP3420 labs/tutorials.

Your report must contain the following information (please use clear section headers, the ones given in boldface below, for each):

(a) **Data Exploration:** A description of the data exploration steps you have done, and what you found out about the data quality of this data set. You must include details about the attributes (their names, types, for example numerical, categorical, ordinal, etc.). You can use a table for this. You should also describe the size of the data set (number of records) and the quality of the data, such as missing values, out of range values, distribution of values (means values, minimum and maximum, histograms, etc.). You should include tables and/or figures in your report (1 mark).

(b) **Data cleaning and transformation:** A description of the data cleaning and transformation steps you have done (or not - in which case describe why no cleaning or transformation was needed) (1 mark).

(c) **Three data mining algorithms used:** A description of the three (or more) data mining algorithms/techniques you have used. This must include a justification why you have chosen the algorithms, the parameter settings you have used (and why), and how you evaluated how good your approach was (2 marks).

(d) **Outcomes of your project:** Here you need to describe (and graphically present) the results of your data mining projects. This can be in the form of cluster graphs, accuracy values, error matrices, or ROC or precision-recall graphs. You also need to describe the results you found, and if they make sense to you (1 mark).

Marking: You will receive up-to one mark for each of parts (a), (b) and (d) and up-to two marks for part (c).

The emphasis in this task is on your analysis, interpretation and justification. If you only provide results, whether numerical or graphical you will not receive a good mark.

If you do not provide the name and source URL of your data set we will penalise you one mark!

The report you have to submit for this final task of the assignment must fit into the overall assignment report (which is 10 pages maximum, with minimum 12 point font), including any graphics and tables, and including any references.