F20/21DL Data Mining and Machine Learning
Coursework 2. Bayesian Networks and Clustering

Handed Out: Thursday 18th October 2018.
Work organisation: (*) individual work on on-line tests 1-2 group work; (**) work in groups of 3 students, the group composition is the same as in CW1.
What must be submitted: For individual work, at least a pass in both Test 1 and Test 2 is a pre-requisite for obtaining an individual mark for the group report. For the group work, a report of maximum 4 sides of A4 (five sides of A4 for Level 11), in PDF format, and accompanying software.
To be 'Handed in': 14:00pm GMT Wednesday 7th November 2018 -- via Vision
Worth: 15% of the marks for the module**.**

The point: data clustering and probabilistic data analysis are all important in data mining and machine learning. This coursework gives you experience with each of these things.

In this coursework you will work with the same data sets as in CW1. If you have not tried this yet in CW1, you are strongly encouraged to try to use either command line of Weka (with Bash or any programming language of choice), or embedded Java programming (see the chapter ``Embedded Machine learning in www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf). This will help you to automate your work, concentrate more on analytical (rather than routine) tasks, and improve your programming skills.

What to do:

Everyone: You must pass Test 1 before starting tasks 1-4, and you must pass Test 2 before starting tasks 5-8. These tests prepare you for the analytical parts of the work. Both tests will be available on Vision. The tests are to be completed individually, and not in groups. Individuals failing the tests will not obtain marks for their group report.

The below steps describe your group work tasks:

1. [*Naïve Bayes Networks*] Repeat steps 4-6 from CW1 using Naïve Bayes Networks, record your results clearly in tables, charts or graphs. Use the best No of features that you found in CW1, and use all other tricks, like the best data set size, best splits between tests, etc.
2. [*Make conclusions*] were there improvements compared to CW1? What differences in algorithm performance did you notice? Do Bayes nets help in better answering questions from item 7 in CW1? Or are they more useful for any other reasons, such as performance tuning? If so, what kind of properties and features of Bayes Nets were helpful?

3. [*Beyond Naïve Bayes: complex Bayesian Network Architectures*] Build two or three Bayes networks of more complex architecture for (a smaller version of) this data set, increasing the number of connections among the nodes. Construct one of them semi-manually (e.g use K2 algorithm and vary the maximum number of parents), and two others – using Weka's algorithms for learning Bayes net construction (e.g. use TAN or Hill Climbing algorithms). Run the experiments described in items 4-6 in CW1 on these new Bayes network architectures. Record, compare and analyse the outputs, in the light of the previous conclusions about the given data.
4. [*Make conclusions*] What kind of new properties and dependencies in the data did you discover by means of using the complex Bayesian Network Architectures? Does it help, and

how, to use Bayes nets that are more sophisticated than Naïve Bayes nets? (You may want to read Chapter 6.7, pages 266-270 and pages 451-454 of the Data Mining textbook by Witten et al. before you do these exercises or https://www.cs.waikato.ac.nz/~remco/weka.bn.pdf.)

5. [*Clustering, k-means*] Cluster the data sets fer2017.arff, fer2017*EmotionX*.arff (apply required filters and/or attribute selections if needed), using the k-means algorithm:

5.1. first excluding the class attribute (use *classes to clusters* evaluation to achieve this). This will emulate the situation when the learning is performed in unsupervised manner.

5.2. then including the class attribute. This will emulate the general data analysis scenario.

6. [*Make conclusions*] about the results, compare with classification results obtained in items (1-2).

7. *[Beyond k-means, tools for computation of optimal number of clusters]* Try different clustering algorithms. Try also to vary the number of clusters manually and then use Weka's facilities to compute the optimal number of clusters. Explore various options in Weka that help to improve clustering results. Use the visualisation tool for clustering to analyse the results.

8. [*Make conclusions*] Make conclusions on the obtained improvements to clustering results. Make sure you understand the various details of Weka's output for different (hard and soft) clustering algorithms when clustering is completed. Use Weka's facilities to test the precision of clustering on this data set. Using your work with Weka as a source, explain all pros and cons of using different clustering algorithms on the given data set. Compare to the results of Bayesian classification on the same data set.

Level 11 only (MSc students and MEng final year students):

9. *[Research Question]* Think about your own research question and/or research problem that may be raised in relation to the given data set, and the topics of Bayesian learning and Clustering. Formulate this question/problem clearly, explain why it is of research value. The problem may be of engineering nature (e.g. how to improve automation or speed of the algorithms), or it may be of exploratory nature (e.g. something about finding interesting properties in data), -- the choice is yours.

10. *[Answer your research question]* Provide a full or preliminary/prototype solution to the problem or question that you have posed. Give logical and technical explanation why your solution is valid and useful.

An Important note:

Before you start completing the above tasks, create folders on your computer to store software you produce, classifiers, Weka settings, screenshots and results of all your experiments. As part of your coursework marking, you may be asked to re-run all your experiments in the lab. So please store all of this data safely in a way that will allow you to re-produce your results on request.

What to Submit::

(a) All evidence of conducted experiments: data sets, scripts, tables comparing the accuracy, screenshots, etc. Supply a link to your HW web space, github or Google drive.

(b) A report of maximum FOUR sides of A4 (11 pt font, margins 2cm on all sides) for Honours BSc students and FIVE sides of A4 (11 pt font, margins 2cm on all sides) for MSc students, containing the following:

Everyone:

HOW: up to a half page each describing how you did steps 1, 3, 5 and 7.

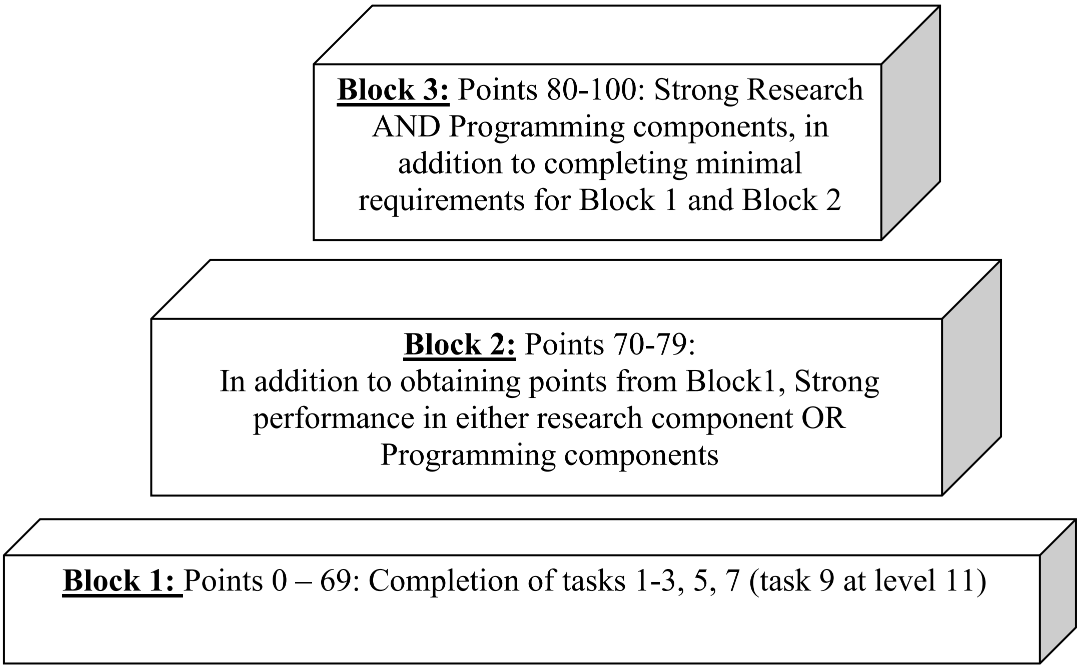
RESULTS: up to two and a half pages of algorithm and data set analysis guided by questions 2,4,6,8.

Level 11 only:
In addition, about a page devoted to tasks 9-10.

Marking: see rubrics on Vision.
Maximum points possible: 100.

You will get up to 69 points (up to B1 grade) for completing the tasks 1-3,5, 7 well and thoroughly (and task 9 for level 11).
In order to get an A grade (70 points and higher), you will need to do well in tasks 4,6,10 by showing substantial skills in either research or programming:

- Research skills: Higher marks will be assigned to submissions that show original thinking and give thorough, logical and technical description of the results that shows mastery of the tools and methods, and understanding of the underlying problems. The student should show an ability to ask his/her own research questions based on the CW material and successfully answer them.
- Programming skills: You will need to produce a sizeable piece of software produced to automate some tasks.
- The mark distribution will thus follow the below scheme:



Block 3: Points 80-100: Strong Research AND Programming components, in addition to completing minimal requirements for Block 1 and Block 2

Block 2: Points 70-79:
In addition to obtaining points from Block1, Strong performance in either research component OR Programming components

Block 1: Points 0 – 69: Completion of tasks 1-3, 5, 7 (task 9 at level 11)