F21DL Data Mining and Machine Learning: Coursework 1

Handed Out: Monday 17th September 2018.
Work organisation: group work, in groups of 3 students.
What must be submitted: A report of maximum 4 sides of A4 (five sides of A4 for Level 11), in PDF format, and accompanying software.
Submission deadline: 15:00pm Tuesday 16th October 2018 -- via Vision
Worth: 15% of the marks for the module**.**

The point: Data preparation and analysis, confusion matrices, correlation and feature selection are all important in real-world machine learning tasks. So this coursework gives you experience with each of these things. You will also experiment with the Nearest Neighbour machine learning method.

In this coursework you will work with a big 'emotion recognition' dataset, created by the research group of Pierre-Luc Carrier and Aaron Courville. Get a link to it from
Vision -> Assessment -> Coursework 1

The data set (of 35887 examples) consists of 48x48 pixel grayscale images of faces. The faces have been automatically registered so that the face is more or less centered and occupies about the same amount of space in each image. The task is to categorize each face based on the emotion shown in the facial expression in one of seven categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral).

In the zip file, the main data set is called fer2018.csv. In it, the first field corresponds to the class field: categories (0=Angry, 1=Disgust, 2=Fear, 3=Happy, 4=Sad, 5=Surprise, 6=Neutral). The remaining fields correspond to a 48x48 array image, so that each field indicates the general amount of ink in a specific area of the image. For example, if in fer2018.csv, the first example has row has first field given by '6', then it means that the first picture shows a face with neutral emotion.

This main file is supplemented by 7 csv files, in which each of the above 7 emotions is tested separately:

- fer2018angry.csv,
- fer2018disgust.csv
- etc.

In all these datasets the class field is either 0 or 1. It is '1' if the image shows the emotion 'X', and it is 0 otherwise. For example, in fer2018angry.csv, if the class field is '1', then that instance corresponds to an image with an angry face; if the last field is '0', then that instance corresponds to a face with any other emotion.

What to do

Everyone:

Form a join the group in which you will work; discuss with the group your strategy for completing the courseworks: the workload split, the tools, the methods… Please use Vision **F21DL 2018-2019: Data Mining and Machine Learning** (subpage Assessment) to register your group or join an existing group.

Choose the software in which to conduct the project. We strongly recommend all students use Weka. Weka is a mature, well-developed tool designed to facilitate mastery of machine-learning algorithms. It is supported by a comprehensive textbook: http://www.cs.waikato.ac.nz/ml/weka/book.html . Weka supports embedded Java programming, and you are welcome to use embedded programming in this assignment as it will allow you to automate parts of this assignment. (See the chapter ``Embedded Machine learning in www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf). Alternatively, the Weka command line interface may be embedded inside Bash (shell) scripts, instead of Java.

Students wishing to complete the below tasks in other languages, such as R, Matlab, Python are welcome to do so, assuming they have prior knowledge of these languages.

In the below task spec, the assumption is made that you are using Weka. Please adapt the below instructions accordingly if you use a different programming language.

After collecting the files as above, you will:

1. *[Data Conversion]* Convert all csv files into arff format suitable for Weka. You should have a Weka data set with 2305 attributes as a result. Some suitable data set pre-processing will be needed before you can load the csv file to Weka.
2. *[Data Randomisation]* Produce versions of these files that have the instances in a randomised order.
3. *[Reducing the size, dealing with computational constraints]* The given files may be too big for standard settings of GUI Weka: decide how you are going to deal with this:

- You may reduce the number of attributes, as taught during the course. Record and explain all choices made when you perform the reduction of attributes. A number of algorithms and options are available in Weka. See Sections 2.1 -2.3 in www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf
- Alternatively, you may use the full data set and the Weka command-line interface. See Section 5 of www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf) and manipulate the heap size (see https://weka.wikispaces.com/OutOfMemoryException).
- Either choice is acceptable as long as you can perform the next task.

4. *[Classification: Performance of the Nearest Neighbour Algorithm on the given data set]* Run the iBK (Nearest Neighbour) tool in Weka on the resulting version of fer2018.arff, with the number of neighbours set to 3. To be able to do this, you may need to apply several Weka "Filters". Explain the reason for choosing and using these filters. Once you can run the algorithm, record, compare and analyse the classifier's accuracy on different classes (as given by the Weka Summary and the confusion matrix).
5. *[Deeper analysis of the data: the data is split into 7 classes, search for important attributes for each class]* For each fer2018*EmotionX*.arff file:

- Using the Weka facility "Select Attributes" for each of these 7 files, record the first 10 fields, in order of the absolute correlation value, for each emotion.

6. *[Try to improve the classification, based on information from 5*] Using the information about the top correlating features obtained in item (5), transform the full data set fer2018.arff so as to keep the following attributes:

- Using only the top 2 non-class fields from each fer2018*EmotionX*.arff.
- Using only the top 5 non-class fields from each fer2018*EmotionX*.arff.
- Using only the top 10 non-class fields from each fer2018*EmotionX*.arff.

- You will have three data sets, with 14, 35 and 70 non-class attributes respectively. Repeat the experiment described in item (4) on these three data sets.

7. *[Make conclusions:]* What kind of information about this data set did you learn, as a result of the above experiments? You should ask questions such as: Which emotions are harder to recognise? Which emotions are most easily confused? Which attributes (fields) are more reliable and which are less reliable in classification of emotions? What was the purpose of Tasks 5 and 6? What would happen if the data sets you used in Tasks 4, 5 and 6 were not randomised? What would happen if there is cross-correlation between the non-class attributes? You will get more marks for more interesting and ``out of the box" questions and answers. Explain your conclusions logically and formally, using the material from the lecture notes and from your own reading to interpret the results that Weka produces.

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

1. *[Research Question]* What are the effects of choosing different numbers of neighbours for the Nearest Neighbour algorithm, instead of the 3 used in the previous tasks? How is this affected by noise levels in the data? Your answer should consider the theoretical possibilities (based on lectures and on your own reading of research papers and textbooks) and some experimental results using the emotion dataset.

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. Archive these folders and submit via Vision. 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

You will 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:
1. HOW: up to a half page each describing how you did steps 1, 2, and 3.
2. RESULTS: up to two and a half pages showing and discussing the results from step 4 to step 7 (I expect these to include a discussion or display of the selected settings, confusion matrices for Nearest Neighbour).
3. Figures and screenshots should take about 1 page.

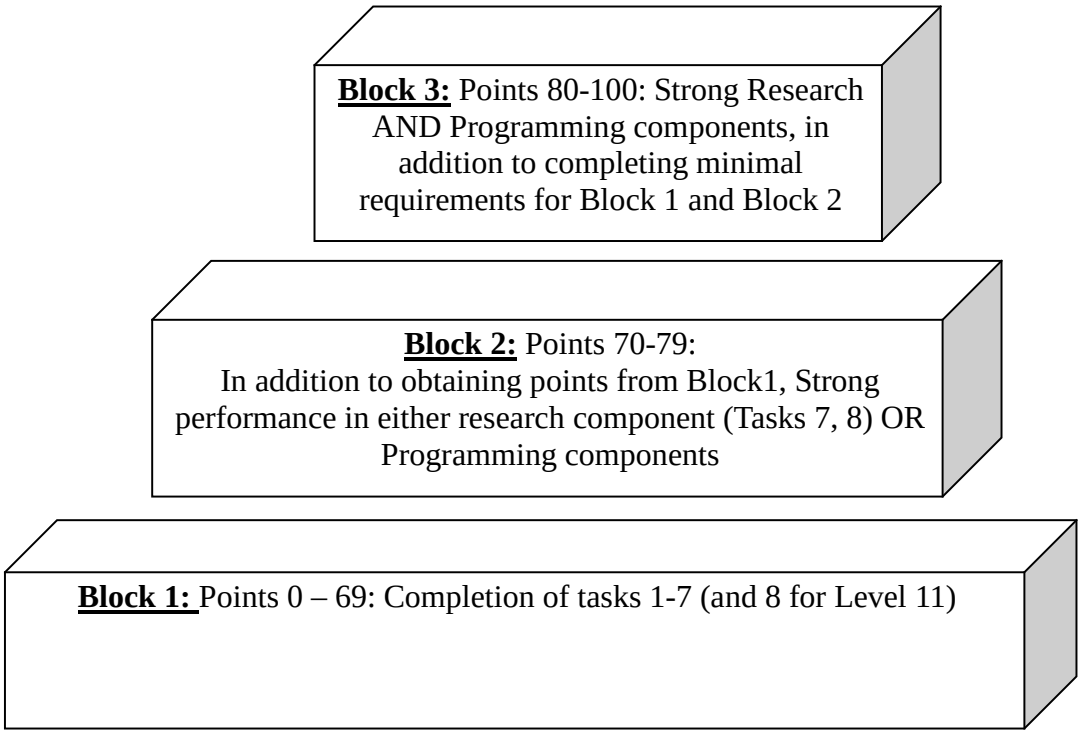
Level 11 only:
In addition, about a page with the title "The effects on Nearest Neighbour of varying numbers of neighbours"

Marking: See Rubric on Vision
Maximum points possible: 100.

You will get up to 69 points (up to B1 grade) for completing the Tasks 1-7 (and Task 8 for Level 11) well and thoroughly.

In order to get an A grade (70 points and higher), you will need to do well in Tasks 1-7 (and 8 for Level 11) and in addition, you will need to show substantial skill in either research or programming:

- Research skills: You will need to complete task 7 and explain the results convincingly. 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, evidence of additional reading, and understanding of the underlying problems. You should show an ability to ask your own research questions based on the CW material and successfully answer them. Level 11 students are also expected to complete Task 8 in depth, using a range of sources and showing you understand in depth the different techniques and the issues in their use.
- Programming skills: You will need to produce a piece of software to cover at least tasks 1-6.
- The mark distribution will thus follow the scheme below:



Plagiarism
This project is assessed as **group work**. You must work within your group and not share work with other groups. Readings, web sources and any other material that you use from sources other than lecture material must be appropriately acknowledged and referenced. Plagiarism in any part of your report will result in referral to the disciplinary committee, which may lead to you losing all marks for this coursework and may have further implications on your degree.
https://www.hw.ac.uk/students/studies/examinations/plagiarism.htm