RMIT UNIVERSITY

Data Mining
COSC 2111/0
Assignment 2

	Assessment Type	You can do this assignment by yourself or in a group of 2. If you are working in a group please establish a group in *Assignment 2 Group* on Canvas. Submit online via Canvas → Assignments → Assignment 2. Marks awarded for meeting requirements as closely as possible. Clarifications/updates may be made via announcements/relevant discussion forums.
	Due Date	End of week 11, Monday 14th October 2019, 9:00am
	Marks	50

1 Overview

In this assignment you are asked to explore the use of neural networks for classification and numeric prediction. You are also asked to carry out a data mining investigation on a real-world data file. You are required to write a report on your findings. You will be assessed on methodology, analysis of results and conclusions.

2 Learning Outcomes

This assessment relates to the following learning outcomes of the course.

- Demonstrate advanced knowledge of data mining concepts and techniques.
- Apply the techniques of clustering, classification, association finding, feature selection and visualisation on real world data
- Apply data mining software and toolkits in a range of applications
- Set up a data mining process for an application, including data preparation, modelling and evaluation

3 Assignment Details

3.1 Part 1: Classification with Neural Networks (15 marks)

This part involves predicting the `Class` attribute in the following file:
`chronic-kidney-disease-2019.arff`
in the directory:
`/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff/UCI/`

The main goal is to achieve the lowest classification error with the lowest amount of overfitting.

For the neural network training runs build a table with the following headings:

Run No	Archi- tecture-	Param eters	Train MSE	Train Error	Epochs	Test MSE	Test Error
1	ii-hh-oo	lr=.2	0.5	30%	500	0.6	40%

1. Describe the data encoding that is required for this task. How many outputs and how many inputs will there be?
2. Develop a script to generate the necessary training, validation and test files. You might want to normalize the numeric attributes with Weka beforehand. Include your data preparation script as an appendix (not part of the page count).
3. Determine the "analyze" strategy that you will use.
4. Using Javanns carry out 5 train and rest runs for a network with 10 hidden nodes. Comment on the variation in the training runs and the degree of overfitting.
5. Experiment with different numbers of hidden nodes. What seems to be the right number of hidden nodes for this problem?
6. For 10 hidden nodes, explore different values of the learning rate. What do you conclude?
7. [Optional] Change the learning function to backprop-momentum. Explore different combinations of learning rate and momentum. What do you conclude?
8. Perform a run with 10 hidden nodes and no validation data. Stop training when the MSE is no longer changing. Get the classification error on the training and test data. Comment on the degree of overfitting.
9. Compare the classification accuracy of the neural classifiers with the classification accuracy of Weka J48 and MultilayerPerceptron.

   Report Length Up to two pages.

3.2 Part 2: Numeric Prediction with Neural Networks (10 marks)

This part involves predicting the `Age` in the following file:
`chronic-kidney-disease-2019.arff`
in the directory:
`/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff/UCI/`

The main goal is to achieve the lowest mean absolute error with the lowest amount of overfitting.

The task is to predict the value of the `Age` variable. Build a similar table of runs to the one in the previous question.

1. Describe the data encoding that is required for this task. How many outputs and how many inputs will there be? What scaling or normalization is required?
2. Modify your script from part 1 to generate the necessary training, validation and test files. You can use Weka to normalize all of the numeric attributes except for the class, ie Age attribute. You will need to write a suitable program to scale the age to the range [0,1] and another one to reverse scale the neural net outputs to get the mean absolute error. Include your data preparation script as an appendix (not part of the page count).
3. Using Javanns carry out 5 train and test runs for a network with 5 hidden nodes. Comment on the variation in the training runs and the degree of overfitting.

   [Hint: When you are comparing the predictive accuracy of different models you don't have to reverse scale the output.]
4. Experiment with different numbers of hidden nodes. What seems to be the right number of hidden nodes for this problem?
5. For 5 hidden nodes, explore different values of the learning rate. What do you conclude?
6. [Optional] Change the learning function to backprop-momentum. Explore different combinations of learning rate and momentum. What do you conclude?
7. Perform a run with 5 hidden nodes and no validation data. Stop training when the MSE is no longer changing. Get the error on the training and test data. Comment on the degree of overfitting.
8. Compare the mean absolute error of the neural classifiers with the mean absolute error of Weka M5P and MultiLayerPerceptron.

Report Length Up to one page.

3.3 Part 3: Data Mining (25 marks)

This part of the assignment is concerned with the files
`portugal-students.arff`
`portugal-students.txt`
which are in the directory:
`/KDrive/SEH/SCSIT/Students/Courses/COSC2111/DataMining/data/arff`

The file `portugal-students.arff` contains data about students in 2 high schools in Portugal. The file `portugal-students.txt` contains a description of the data.

Your task is to analyse this data with appropriate classification, clustering, association finding, attribute selection and visualisation techniques selected from the Weka menus and identify any "golden nuggets" in the data. If you don't use any of the above techniques, you need to say why.

Submit: Up to two pages that describe the strategy you adopted, your methodology, the runs you performed, any "golden nuggets" you found and your conclusions.

4 Submission

Submit one pdf file.

After the due date, you will have 5 business days to submit your assignment as a late submission. Late submissions will incur a penalty of 10% per day. After these five days, Canvas will be closed and you will lose ALL the assignment marks.

Assessment declaration:
When you submit work electronically, you agree to the assessment declaration - `https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/assessment-declaration`

5 Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the following: `https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity`.

6 Marking guidelines

Factors contributing to the final mark will include the number of tasks attempted, the amount of exploration of the algorithms, methodology, logical analysis and presentation of results and conclusions.