F21DL Data Mining and Machine Learning: CW1. Your DM&ML Portfolio

Handed Out: Monday 13th September 2021, via Canvas.
Submission deadline: Assessed labs throughout the term will bring intermediate marks, final cut-off date for Portfolio submission is **week 11 (by 27th November 2021 midnight).**
Work organisation: Completed individually during the term. This CW spec also constitutes your Lab spec, i.e. there will be no additional lab exercises. You will be assigned a lab tutor with whom you will discuss your progress during the lab each week.
What must be submitted: Your presentation to the assigned lab tutor will constitute the CW submission. Additionally, at the end of the term, you will submit a Jupyter Notebook on the HWU Gitlab, containing all your code and analysis. This will be needed for audit purposes and for plagiarism checks.
Mode of assessment: The portfolio will be checked by your assigned lab tutor 4 times during the term.
Worth total of: 20% of the marks for the module. Calculated as 20 points, assigned for different parts of the portfolio (as described below).

This coursework is designed to give you experience with, and hence improve your understanding of:

1. Methods for data preparation and analysis, including probabilistic/Bayesian methods of data analysis, calculating correlation of features and performing feature selection.
2. Unsupervised Learning and Clustering
3. Supervised Learning and the problems of generalization and overfitting; Supervised learning methods including Naïve Bayes, Linear Regression, K-nearest neighbors and Decision trees
4. Cutting Edge machine learning techniques: Neural Networks and Convolutional Neural networks

The data set:
This coursework will constitute your own Data Mining and Machine learning portfolio: therefore, the choice of the data set is yours! We suggest that you choose a domain that interests you most. For example, if you are interested in computer games, why not finding or generating your own data set that analyses the player behaviour! When choosing your favourite data set, you may want to take into consideration the following factors:

- Data sets that contain images will require you to learn some basics of image processing, and will be particularly good for working with Convolutional neural nets at the end of the term.
- Data sets that contain nominal (non-numeric) data will work more impressively with Bayes nets and Decision trees than computer vision data sets.
- There are some well-known benchmark data sets, such as Iris, MNIST, CIFAR10, CIFAR100 (an overview will be given in the lectures). We ask that you do not take benchmark data sets, as their properties are already well-explained on-line.
- When searching for data sets, https://www.kaggle.com/ and UCI Machine Learning Repository are good sources.
- You can create your own data set, and this effort will score marks;
- Generally, more challenging data sets will be more likely to require more advanced DM&ML methods, and for this reason may bring higher marks.

Note: you will be asked to use the same data set throughout the whole of your portfolio. So, invest a bit of time into this decision. Ask lab tutors if in doubt.

What to do:

Part 1. Data Analysis and Bayes Nets. To be completed in Week 1-4, and to be presented to your lab tutor no later than Week 5, no marks will be given beyond that week. The interview with the lab tutor will bring up to 5 points on the BSc level, and up to 4 points on the MSc/MEng level.

- Explain your reasons for choosing your data set. If you created the data set yourself, explain all technical difficulties and pitfalls in generating the data set. This part of the assignment assumes that you only have a training set (no testing set). So, all algorithms should only be evaluated on this one data set.
- Visualization and initial data exploration help to gain insights on the data attributes and guides in choosing suitable features and building appropriate ML models. Examine your data through visualization and analysis and show how this helped you learn more about your data and has guided you for further analysis. Discuss how you fixed problems like missing values, errors or outliers -if applicable. Did you need to apply any preprocessing or normalization procedures? If so, why?
- Run Naïve Bayes Classifier on your chosen data set, and record the major metrics: accuracy, TP rate, FP rate, precision, recall, F measure, the ROC area etc (as explained in the lectures). Make conclusions
- Using the methods explained in lectures and tutorials (or additional sources), analyse most correlating features/attributes of the data set, generally and per class. Form 3 data sets, that contain progressively fewer features/attributes.

Example: *suppose my data set has 300 input features and 3 classes. I can find 2, 5, 10 features that best correlate with class 1, 2 and 3 respectively. E.g., for class 1, features 40 and 50 are most correlating. For class 2, features 1 and 5 are most correlating, and for class 3, features 200 and 300 are most correlating. As a result, I will get 3 data sets:*
*Data set 1: contains 2 top features for each of 3 classes: 2*3 = 6 features*
*Data set 2: contains 5 top features for each of 3 classes: 5*3 = 15 features*
*Data set 3: contains 10 top features for each of 3 classes: 10*3 = 30 features*

- Run Naïve Bayes classifier on the resulting 3 data sets, again noticing all major performance metrics.
- Make conclusions: You may want to think about the following questions: what kind of information about this data set did you learn, as a result of the above experiments? Are classes represented equally? Which features are more important/reliable for which class? Which are less reliable? You will get more marks for more interesting and ``out of the box" questions and answers.
- (Optional) You may try to investigate libraries for more complex (non-naive) Bayes nets, repeat the experiments above, and use Bayes nets structure for further feature analysis.

Part 2. Clustering. To be completed in Week 7-8, and to be presented to your lab tutor no later than Week 9, no marks will be given beyond that week. The interview with the lab tutor will bring up to 5 points on the BSc level, and up to 4 points on the MSc/MEng level.

- Using the same data set, use k-means clustering to find clusters in your data set. Evaluate the accuracy of this clustering, visualize the clusters, make conclusions.
- For top marks, try different clustering algorithms for hard and soft clustering, such as EM, GMM, hierarchical clustering or any other algorithms of your choice. Compare their performance on your data set, make conclusions.
- Try also to vary the number of clusters manually and then research some of the existing algorithms to compute the optimal number of clusters. How does it affect the accuracy of clustering? Make conclusions.

- (Optional) Look up methods to determine the optimal number of clusters. For example, look up: Elbow method, the silhouette method, cluster validity and similarity measures.
- Using your experiments as a source, explain all pros and cons of using different clustering algorithms on the given data set. Compare the results of Bayesian classification on the same data set.

Part 3. Supervised Learning: Generalisation & Overfitting; Decision trees. To be completed in Week 8-9, and to be presented to your lab tutor no later than Week 10, no marks will be given beyond that week. The interview with the lab tutor will bring up to 5 points on the BSc level, and up to 4 points on the MSc/MEng level.

- Make sure that you obtained or created a test set.
- Use Decision trees (the J48 algorithm) on a training set, measure the accuracy. Then measure the accuracy on the training set using 10-fold cross-validation. Record all your findings and explain them. Use the major metrics: accuracy, TP rate, FP rate, precision, recall, F measure, the ROC area if needed.
- Repeat the experiment, this time using training and testing data sets instead of the cross validation. That is, build the J48 classifier using the training data set, and test the classifier using the test data set. Note the accuracy. Answer the question: Does the decision tree generalize well to new data? How do you tell?
- Experiment with various decision tree parameters that control the size of the tree. For example: depth of the tree, confidence threshold for pruning, splitting criteria and the minimal number of instances permissible per leaf. Make conclusions about their influence on the classifier's performance.
- Make new training and testing sets, by moving 30% of the instances from the original training set into the testing set. Note the accuracies on the training and the testing sets
- Make new training and testing sets, by moving 60% of the instances from the original training set into the testing set. Note the accuracies on the training and the testing sets
- Analyse your results from the point of view of the problem of classifier over-fitting. Do you notice the effects of over-fitting? How? Note your conclusions in the Jupyter notebook.
- For higher marks, try some other decision tree algorithms (e.g. random forests). Repeat all of the above experiments and make conclusions.

Part 4. Neural Networks and Convolutional Neural Networks. To be completed in Week 9-10, and to be presented to your lab tutor no later than Week 11. The interview with the lab tutor will bring up to 5 points on the BSc level, and up to 4 points on the MSc/MEng level.

In this part, you will use the original training and testing data sets.

- Run a Linear classifier on the training data set, with 10-fold cross validation and without, mark the accuracies. Note also its accuracy on the test set. How well does the linear classifier generalize to new data? What hypothesis can you make about this data set being linearly separable or not?
- Run the *Multilayer Perceptron,* experiment with various Neural Network parameters: modify the activation functions, experiment with the number and size of its layers, vary the learning rate, epochs and momentum, and validation threshold. Analyse relative performance of the resulting Neural Networks and changing parameters, using the training and the test data.
- Based on all of these experiments, what conclusions can you make about the data set complexity (linear separability), and the capacity of deep neural networks to generalize to new data? Can you make any conclusions about the effect of activation functions?
- For top marks, repeat these experiments using Convolutional Neural networks. For this types of networks, you can additionally vary the kinds of layers (convolutional, pooling, fully connected).

Part 5. Level 11 only (MSc students and MEng final year students). To be completed and presented any time during the term, but no later than Week 11. The interview with the lab tutor will bring up to 4 points on the MSc level.

[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 completed portfolio. 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.
[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.
Please start thinking and researching Part 5 (Research question) early in the course. Don't leave it until the last weeks. Week 6 is a good timing to start.

Final remarks:

Plagiarism
This project is assessed as individual work. You must not share your work with other students. 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 coding or data/algorithm analysis 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

Lateness penalties
Standard university rules and penalties for late coursework submission will apply to all coursework submissions. See the student handbook.