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

Handed Out: Monday 25th September 2023, via Canvas.
Submission deadline: **Please consult submission deadlines for each coursework part below**
Work organisation: Work in groups of (3-5). You will receive weekly suggested lab tasks to complete individually and with your group. The tasks will help you build up your portfolio below. You can ask your tutor/instructor for feedback and recommendations during weekly lab hours.
What must be submitted: Submit a link to Jupyter Notebook, containing all your code and analysis. Submit a brief report 1-2 pages summarizing your findings and insights.
Mode of assessment: Marking criteria will include quality of code submitted, depth of analysis and discussions and engagement during lab hours. Marking rubric is provided as a general guideline.
Worth total of: 40% of the marks for the module.

This coursework is designed to help you understand and enhance your experience with:

1. Methods for data exploration, preparation and analysis, including 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:

In your coursework you will work with an image data set (Traffic sign data set). A detailed description of the data set together with all downloadable files can be found under your assessment link. You will report your main findings on this data set but feel free to experiment and run your code with other data sets as well.

Exploring different data sets might be useful for future projects, for your dissertation or if you would like to score bonus points by showing "extra effort and creativity". Starting your experiments with a simple data set can help you understand some basic concepts.

What to do:

Part 1. Data Analysis and Bayes Nets. To be completed in Weeks 1-5

This part of the assignment assumes that you are using a training set (keep your testing set aside for now). So, all algorithms should only be evaluated on the training 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. Use cross-validation on your training set to report your finding. Alternatively, you can generate a stratified train-test split version.

- Using the methods explained in lectures and tutorials (or additional sources), analyze most correlating features/attributes of the data set, generally and per class. Form 3 data sets, that contain progressively fewer features/attributes.

For your coursework data set, try to create 3 data sets with 50, 100 and 200 features (pixels) each. Choose the top correlating pixels/class (use the *OnevsAll* training data sets).

In particular, for each of the 10 classes find the *5, 10, 20 features(pixels) that best correlate with classes(0-9).*

As a result, you will get 3 data sets:

*Data set 1: contains 5 top features for each of 10 classes: 5*10 = 50 features*

*Data set 2: contains 10 top features for each of 10 classes: 10*10 = 100 features*

*Data set 3: contains 20 top features for each of 10 classes: 10*3 = 200 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 the 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 Weeks 7-9.

- For the same data set, use k-means clustering to find clusters in your data. 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 Weeks 8-10.

- Now you will start working with the provided test data sets.
- Use Decision trees (the J48 algorithm) on the 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 Weeks 9-11.

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. Analyze relative performance of the resulting Neural Networks and changing parameters, using the training and the test data. What techniques can be used for performing hyperparameter tuning in a systematic way? Report on the best combination of parameters obtained for your experiments.
- 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 these 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 no later than Week 11.

[Research Question] Think about your own research question and/or research problem that may be raised in relation to the given data set, and your portfolio tasks. 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. You can also try to solve problems marked as 'optional' or 'for higher marks'. Those are good candidates to add to your research question.
[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.

When to submit: (You should complete your portfolio tasks by the following dates)

- Part1(15%): Data Analysis and Exploration - 23.10.2023 by 8 am
- Part2(20%): Clustering - 06.11.2023 by 8 am
- Part3(25%): Decision Trees -13.11.2023 by 8am
- Part4(25%): Neural Networks and CNN - 27.11.2023 by 8 am
- Part5(15%)(PG) only: Research question - 27.11.2023 by 8am

What must be submitted: (per group) - you will submit the following for each part separately

- Link to a well documented Jupyter Notebook, containing all your code and analysis.
- 1-2 page report summarizing your findings and insights. Figures and tables do not count towards the page limit. Report should include declaration of what parts of the coursework each group member contributed. Data preparation, programming, analysis, report writing (and generation of figures and illustrations for the report) all count as contribution.
- Group authorship declaration form required by the University.

Marking criteria: (see marking rubric for guidance)

Evaluating your work will be based on:

- quality of notebook provided
- demo of code in lab[1] and depth of discussions
- analysis of results and graphs and insights provided (included in report and shown in notebook and during lab discussions)
- level of engagement in group work.

In addition, 10% individual marks will be assigned. No marks for non-contributing group members

Recommendations for Group work:

-Each student is advised to keep his/her own copy of the python notebook and run additional experiments if they wish (maybe you have an extra data set you want to experiment with?!)
-Group members can communicate virtually to complete the weekly portfolio tasks. Please minute your weekly discussions and upload it to your Github, Bitbucket, group space on canvas (or as agreed by your instructor/tutor)
-Attendance of groups each week in the lab is mandatory
- Please assign a group rep to communicate with your tutor/instructor if needed

How to make the best of your group:

-Divide tasks and do better research on each topic
-Share experiences and explore wider ideas. Conduct more experiments.
-Discuss your findings among group members. Try to explain and justify your findings.

Plagiarism

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.

[1] Details of demo schedule will be communicated in your campus by your instructor.