F21DL Data Mining and Machine Learning: Your DM&ML Portfolio, April Resit.

Handed Out: Wednesday 15th June 2023, via Canvas.

Submission deadline: Final cut-off date for Portfolio submission/discussion is **27th July 2023 midnight.**

Work organisation: Work is done individually; you are allowed to use the code given within the Python tutorials on F20/21DL Canvas pages.

What must be submitted: Submit a Jupyter Notebook, containing all your code, experiments, discussion and analysis via Canvas. This will be needed for audit purposes and for plagiarism checks.

Mode of assessment: By interview arranged as per announced schedule. Check more details on canvas page. Each interview will last 30 minutes, you must ensure that you can show your Jupyter Notebook during the interview.

Marking criteria will include quality of code submitted, depth of analysis and discussion, demonstrated expertise in the subject at the interview. Each of the four sections below will bear 25 points each; in each section, you will get up to 13 points for obtaining the basic functionality and basic experiments, and up to further 12 points – for parts that are marked as "***For Top Marks***" 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:

You will work with the data set that contains happy, sad, and neutral faces.

On Canvas, you will find two versions of data, with noise and without.

For top marks, Noisy images may help you to make more interesting arguments in some experiments.

What to do:

Part 1. Data Analysis and Bayes Nets.

- 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 (in the Jupyter Notebook).
- **For top marks:** 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 2 data sets, that contain progressively fewer features/attributes.

Example: In your data set, you can find 3 and 6 features that best correlate with class 1, 2 and 3, respectively.

As a result, you will get 2 data sets:
*Data set 1: contains 3 top features for each of 3 classes: 3*3 = 9 features*
Data set 2: contains 6 top features for each of 3 classes: 6 3 = 18 features*

Run Naïve Bayes classifier on the resulting 2 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.

Part 2. Clustering

- 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.
- **For top marks,** 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.

- Create 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.
- **For top marks:** 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. Then once again, 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 top 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.

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). Make conclusions about performance of these networks compared to other machine learning algorithms.

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.