DALT7012
Advanced Machine Learning
Semester 2, 2021/22
Coursework Assignment

OXFORD BROOKES UNIVERSITY

Learning Outcomes

This coursework is worth 100% of your module marks and provides the following module-level outcomes.

1. Apply and evaluate complex machine learning algorithms for use in high dimensionality, multi-modal environments.
2. Critically apply a theoretical analysis of a machine learning algorithm to rank its appropriateness and limits in a particular scenario.
3. Create solutions to complex, unbounded machine learning problems using appropriate software.

Submission Guidelines

The report should be submitting via Turnitin from the Module's Moodle web site as a single file (preferably in PDF).

Requests for extensions should be made via the Extenuating Circumstances procedure as usual.

Academic Conduct Regulations and Procedure:
This is an individual assignment, therefore evidence of academic misconduct, such as group work or collusion, will be penalised. For more information see:
https://www.brookes.ac.uk/students/sirt/student-conduct/academic-misconduct/

Any use of the work of other people, including code and algorithms, must be clearly referenced. See the regulations for details: https://www.brookes.ac.uk/regulations/

Assessment and Feedback

There are two feedback opportunities and two submission points.

The teaching sessions in week 4 will be used as the first opportunity to discuss your plans for the coursework and prepare initial submission, so it is important that you follow the lectures in the first few weeks and think immediately about how to apply the material you have learned.

You are expected to have made substantial progress on points (1)-(3) by **week 8**, and you can submit a preliminary report on these points to receive formative feedback from the tutor. All initial submissions will receive feedback on Friday of week 9. This is a formative submission, therefore no marks will be awarded, just suggestions on how to improve your report.

Submit your final report in **week 12**, digitally as a single file on Moodle. The final report should be a final revised version of the initial submission, taking the comments into consideration.

Both marks and final feedback will be given after the exam committee.

Deadlines

Coursework released	**Friday of Week 4**, 18th February 2022
Initial Submission Deadline (optional)	Friday of Week 8, 18th March 2022.
Formative Feedback	Friday of Week 9, 25th March 2022.
Final Submission Deadline	**Monday of Week 12**, 25th April 2022, 17:00:00.

Introduction

Recognising facial expressions is a difficult problem, and there is lots of research in this area. In particular, Grammatical Facial Expressions are used in sign language, and carry a specific meaning ('affirmative', 'negative', and so on). Grammatical Facial Expressions are often used to qualify specific hand signs or to complete the emotion related to the signs.

The goal of this assignment is, having selected one of the 9 available expressions (e.g., 'affirmative'), to classify each test frame, represented by a vector with 300 components, collecting the *x,y,z* coordinates of the 100 landmarks making up the facial expression, as either a positive example (i.e., the expression there is affirmative) or a negative example (the expression there is *not* affirmative).

Hence, this is a **binary classification problem**: either the desired expression is present, or it is not present.

For this coursework, you are required to apply two machine learning methods to a specific data set. This data set represents coordinates of landmarks outlining the contours of faces, each extracted from an individual frame of a video. The dataset collects videos which show people performing basic facial expressions. The task is, after training a classifier on a collection of training frames, to categorise each test frame as either a positive (i.e., that expression is present in that frame) or a negative (that expression is not present in that frame) example of a specific expression. The data was used in a research paper, which is attached on Moodle. It is advisable to read the paper to better understand the data set and the task.

- Read the entire coursework specification sheet thoroughly.
- Read the dataset description and further details on the data (see Appendix B).
- Read the paper carefully, and ask questions of the lecturer or tutor if anything is not clear.

All experiments must be thoroughly explained and accompanied with graphs and plots, where appropriate, and a set of R scripts from which they can be reproduced.

Dataset

The dataset proposed in this Coursework has been already used in the experiments described in the paper by Freitas et al. (2014), listed on Moodle. The dataset is composed by eighteen videos recorded using Microsoft Kinect. In each video, a user performs (five times), in front of the sensor, five sentences in Libras (Brazilian Sign Language) that require the use of a grammatical facial expression. The dataset considers 9 such Grammatical Facial Expressions (GFEs): *affirmative, conditional, doubt question, emphasis, negative, relative, topics, wh-question* and *yn-question*.

Each video sequence in the dataset is thus a series of frames portraying each a GFE. By using Microsoft Kinect, the scientists obtained: (a) an image of each frame, identified by a timestamp; (b) a text file containing the coordinates (x, y, z) of 100 landmarks outlining the contours of the facial expression (see Figure 1) from eyes, nose, eyebrows, face contour and iris.

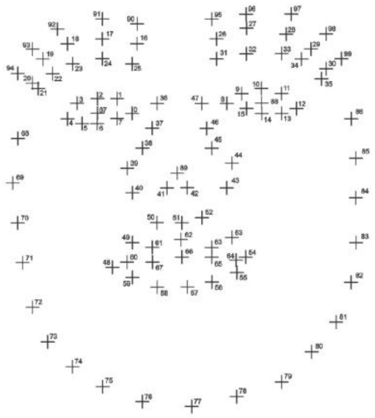

Figure 1: Location in the image plane of the 100 landmarks, which describe an example facial expression from the dataset.

There is one such text file per expression, for each user (A or B). So, in total, 18 text files. In all such files, each line collects the coordinates of the 100 landmark points extracted from a single frame. Other 18 files contain the class (positive or negative) of each frame. The name of the file refers to each video: the letter corresponding to the user (A and B), name of grammatical facial expression and a specification (target or data points). The images enabled the manual labelling of each file by a specialist, providing the necessary ground truth for classification.

The **Grammatical facial expressions dataset** can be downloaded from the UCI Machine Learning Repository:

https://archive.ics.uci.edu/ml/datasets/Grammatical+Facial+Expressions

Dataset creators: Fernando de Almeida Freitas, Felipe VenÃ¢ncio Barbosa, Sarajane Marques Peres.

The authors request a citation to the paper: Freitas, F.A., Peres, S.M., Lima, C.A.M., Barbosa, F.V. (2014). 'Grammatical Facial Expressions Recognition with Machine Learning', in: *27th Florida Artificial Intelligence Research Society Conference (FLAIRS), 2014, Pensacola Beach. Proceedings of the 27th Florida Artificial Intelligence Research Society Conference (FLAIRS).* Palo Alto: The AAAI Press, pp. 180-185.

Suggested Contents, Tasks and Marking Scheme:

Title page

The following information should appear on the front page of your report:

- Module number: DALT7012.
- Student Number:
- MSc Course: MSc in …
- Word count:

1. **Train the classifier on expressions from user A (22%)**

Use an appropriate classification method among those you know (e.g. hard SVM, soft SVM, K-NN, CNN, etc.) to train the classifier on expressions from user A, broken down into:

   a. Basic implementation, using an off-the-shelf classifier (e.g. a standard R command), on a single facial expression [10%];

   b. Repeat the training for a different facial expression [4%];

   c. Extra marks for coding your own implementation of the chosen classifier [8%].

2. **Evaluate accuracy (20%)**

Employ this trained classifier on the expressions performed by user B, and determine the accuracy, broken down into:

   a. Test on a single facial expression, using an off-the-shelf classifier, and comment on the results [10%];

   b. Test on a different facial expression, and comment on the difference [5%];

   c. Use of performance measures other than simple accuracy (e.g. precision, recall, Fscore, ROC curve) [5%]. A useful link is https://www.cs.cornell.edu/courses/cs578/2003fa/performance_measures.pdf

3. **Additional experiments on the first classification method (8%)**

   a. Repeat the analysis (training + testing) by inverting the roles of the user (i.e., training on user B and testing on user A), and comment on the difference [4%];

   b. Repeat the analysis (training + testing) by considering a different feature representation than the original landmark coordinate vector, and comment [4%]. For instance, you can use the two representations proposed in the paper by Freitas et al, you can try PCA or any other dimensionality reduction, and so on.

4. **Second ML Technique and Performance Evaluation (30%)**

Choose a different classifier than before, and repeat the training and testing, broken down into:

   a. Basic training of an off-the-shelf classifier on a single facial expression [8%];

   b. Testing of an off-the-shelf classifier on a single facial expression [4%], commenting on the results;

   c. Extra marks for own implementation of the second classifier [3%];

   d. Training the second classifier on an extra expression [3%];

   e. Testing the second classifier on an extra expression, with comments [3%];

   f. Repeat the analysis (training + testing) by inverting the roles of the user (i.e., training on user B and testing on user A), and comment on the results [3%];

   g. Repeat the analysis (training + testing) by considering a different feature representation than the original landmark coordinate vector, and comment on the results [3%].

   h. Use of performance measures other than simple accuracy [3%].

5. **Conclusion (20%)**

Compare the results of the two classifiers, and comment on their performance. Summarise your main findings. Explain possible current limitations of your solutions and possible further strategies to improve on the results. All arguments must be evidence-based.

References

Provide a list of references.

You are required to cite the work of others used in your solution, include a list of references, and avoid plagiarism and collusion. Remember that each loan should have at least one citation (use the university recommended referencing style).

Appendix

Append your original source code (e.g., R code) in case you decide to build your own classifier or add the description of the packages and commands employed otherwise. Please do not attach your code to the report for off-the-shelf classifiers, but send it to the Module Leader via email before the deadline.

References and Appendices themselves will not be marked. However, inappropriate usage of these sections or their absence will be taken into consideration when awarding the final mark (see Marking Rubric).

The number of marks for each Task and Subtasks are the maximum achievable.

Recommended software

We are **agnostic about your choice of implementation** (although working in R would be a natural choice).

Report format

All above mentioned components should be included in a single file (preferably in PDF).

The assignment must be presented in the following format:

- Font must be 11 point Arial font, Line Spacing – single and Spacing After – 10 pt.
- All pages must be numbered.
- Margins must be as follows: Top: 1 inch, Bottom: 1 inch (2.5 cm), Left: 1.25 inches, Right: 1.25 inches (3.2 cm).

Report word limit: **2,000** words.

Word count excludes cover sheet, title, tables, figure labels, bibliography and appendices.

Words that exceed the maximum stated words (+10%) will not be marked. If in doubt, you should discuss this with the Module Leader **before** submission.

Final provisions and recommendations

There is no 'correct' choice of classifiers, as what performs better on this task usually depends on the specific expression you choose. Also, marks do not depend on who will get the best performance (since you might work on different expressions, it would be impossible to find out anyway), but on the depth of your analysis and the interest of your solution.

Note: Training a classifier
Almost every classifier depends on a number of parameters (for instance, for neural nets the parameter p is the number of layers and the number of neurons per layer; for support vector machines, p is the parameter C that weighs the importance of the slack variables, and so on).

Therefore, for training purposes (Tasks (a), (c) and (d)), you need to find the value of the parameter vector which delivers the best performance. Standard practice is to do this by cross-validation: https://en.wikipedia.org/wiki/Cross-validation_(statistics).

Cross-validation amounts to:

a. Dividing the training set (e.g. all the ”emphasis” expressions for user A) into K validation folds (the choice of K is up to you - do you get different results for different values of K?)
b. Exploring the possible values of your parameter(s) and for each value:
   (i) training your classifier with that parameter value on $K - 1$ folds, and testing it on the remaining fold, for all possible choices of the test fold;
   (ii) this will give you a performance number (e.g., accuracy) for each fold, for the current value of the parameter;
   (iii) taking the average over all the K folds - that yields the training performance of your classifier for the current parameter value;
c. Pick the parameter value with the highest reported average performance.

Appendix A: Marking Rubric

Criteria	F 0-39%	D/C 55%	B 65%	A 75%	Weight
Training the first classifier from user A	Poor implementation of the first classifier. Many mistakes and discrepancies. Quite unclear and brief description of the experiment. No results on user A or they are confusing.	Adequate implementation, using an off-the-shelf classifier, but the model is not described properly (e.g. what model is implemented is unclear). Limited information related to the first classifier and weak interpretation of results and details on cross-validation.	Good data preparation and implementation of the first classifier but some details on the model are missed or the model and its hyperparameters are not thoroughly explained.	Excellent data preparation with own implementation of the first classifier. Pre-processing was correctly applied and the results of cross-validation were discussed in detail. Results are accompanied by proper visualisation and statistics. All in a clear, concise and comprehensive manner.	**22**
Testing the first classifier on user B	Tests are incomplete or incorrect and almost nothing in terms of commenting on the results. There is an attempt to use of performance measures other than simple accuracy but no proper comment or they are confusing. No interpretation of the results.	Adequate test implementation and additional performance metrics are used but not described properly. No mention performance on the training dataset. Fair discussion on the results but details are missing. Not a lot of interpretation.	Good testing for both facial expressions. Good use of other performance metrics, but not too much in terms of their interpretation or they are not described in details. Interesting analysis but not too much in detail.	Excellent testing and commenting on the results. All experiments are thoroughly explained. Appropriate original diagrams and tables, with captions are provided and commented properly. All statistical metrics and their values (e.g. precision, recall, F-score, ROC curve) are explained and analysed deeply.	**20**
Additional experiments	Poor experiment realisation. Some results have been provided but without trying to interpret them. Missing information about the impact of feature representation on the first classifier. Illustrations / plots are swapped or not present.	Acceptable realisation of experiments with (training + testing) by inverting the roles of the user but commenting on the difference are weak or absent. The analysis by considering a different feature representation than the original landmark coordinate vector exists but there is there is no proper explanation on how does it affects on the performance of the first classifier.	Good implementation of the experiments by inverting the roles of the user and good commenting on the difference. Some minor details may be absent. Good explanation on effect of different feature representation. Good presentation with original diagrams or tables, and proper commenting and discussion, but discussion somewhere needs improvement.	Excellent realisation of the experiments by inverting the roles of the user and commenting on the difference. A clear motivation on choosing different feature representation. Excellent implementation, presentation and discussion. Clear & logical argumentation and analysis.	**8**
Second ML technique and	Poor implementation of the second classifier. Many subtasks are missing or realised with errors. Quite unclear or too brief description	Correct implementation of the experiments, but the second model is not described properly or comments related to the second classifier are limited. Weak interpretation of results,	Good implementation of the second classifier and good presentation, but some details on the model are missed or the model and its parameters are not thoroughly explained (e.g. weak	Excellent implementation of the second classifier for two facial expressions. Results are accompanied by proper visualisation and statistics. A set of	**30**

performance evaluation	of the experiments. No results were provided or they are confusing.	no evidence of using cross-validation or just listing the results. Some additional statistical metrics are provided but they are not explained in detail.	argumentation on choosing hyper parameters). All statistical metrics and their values (e.g. precision, recall, ROC curve) are justified, but discussion needs improvement.	additional statistical metrics and their values are provided, explained and analysed deeply. Clear & logical argumentation and analysis.	
Conclusion	Conclusions are inadequate or missing. Poor critical analysis, e.g. the analysis is inadequately short, very limited interpretation of the results, etc.	Acceptable analysis. Some interesting statements and comparisons, but discussion is quite short. The limitations of the solutions may be not considered. The possible further strategies to improve the results may need further elaboration. The references are provided.	Good conclusion, drawing on the source materials. Strong evidence of critical analysis. The limitations of your solutions are discussed but not much on further strategies. Some of the statements are a bit vague or imprecise. The analysis could have been pushed a bit further in terms of data.	Excellent analysis with good comments and deep insight. Pervasive and original thoughts and vision of the problem. The discussion is extensive and interesting. The performance and drawbacks of the approaches were explored and evaluated very carefully. Outstanding representation of further strategies.	**20**

Appendix B: Further details on the data

Number of frames per expression, per user

Number of frames: 27965.

As mentioned above, in each frame either user A or user B are portrayed performing one of the 9 Grammatical Facial Expressions described.

Below, (+) frames are those in which the selected expression is present (the positive class of the binary classification task); (-) frames are those in which the expression is not occurring (the negative class in a binary classification task).

The numbers of positive and negative examples of each expression for user A are:

a. a affirmative: 414 (+) frames and 648 (-) frames;
b. a conditional: 548 (+) frames and 1359 (-) frames;
c. a doubt question: 491 (+) frames and 821 (-) frames;
d. a emphasis: 330 (+) frames and 1073 (-) frames;
e. a negative: 528 (+) frames and 596 (-) frames;
f. a relative: 644 (+) frames and 1686 (-) frames;
g. a topics: 360 (+) frames and 1436 (-) frames;
h. a wh question: 609 (+) frames and 677 (-) frames;
i. a yn question: 532 (+) frames and 858 (-) frames.

The numbers of positive and negative examples of each expression for user B are:

a. b affirmative: 528 (+) frames and 546 (-) frames;
b. b conditional: 589 (+) frames and 1445 (-) frames;
c. b doubt question: 780 (+) frames and 717 (-) frames;
d. b emphasis: 531 (+) frames and 813 (-) frames;
e. b negative: 712 (+) frames and 870 (-) frames;
f. b relative: 550 (+) frames and 1354 (-) frames;
g. b topics: 467 (+) frames and 1358 (-) frames;
h. b wh question: 549 (+) frames and 779 (-) frames;
i. b yn question: 715 (+) frames and 1023 (-) frames.

Attributes per frame

- Datapoints files (* datapoints.txt): a timestamp (double) and 100 numeric attributes (double) for each frame;
- Targets files (* targets.txt): a class attribute 0 or 1 (integer) for each frame.

Each landmark point has 3 coordinates: x, y (its location in the image plane, in pixels) and z (its distance from the Kinect camera, in millimetres).