COMS 3007: Machine Learning Assignment 2021

You have, and will still, learn about several classification algorithms in this course. For this assignment, find a suitable dataset on which you can apply various classification algorithms. You may look here for some options: https://archive.ics.uci.edu/ml/datasets.html (but extra credit will be given for a more original data set). This MUST be a classification problem.
Apply X different supervised learning algorithms to your dataset, where

$$X = max\{2, numberOfGroupMembers - 1\}.$$

Extra marks will be given for more than this.

You must **submit a PDF document as well as your code** to Moodle containing the following information. Note that marks are awarded for all the points listed below.

(1) A description of your dataset: where did the dataset come from, what are the attributes, what are the targets, how many datapoints do you have, and give some sample datapoints from the dataset. State what you are trying to predict with the data. How balanced are the classes? Why is this an interesting or important problem that you are looking into? Do not just list the attributes and targets, but discuss what they are and represent.
   Note: marks will be given for an interesting choice of dataset.

(2) A description of how you structured your inputs/targets and normalised and preprocessed the data. What did you do with missing values, strings, etc? In addition, how did you split into training/validation/test data. Motivate your choices for all of these.

(3) A list of classification algorithms you applied to the data, together with the details of each implementation. Why did you choose these algorithms? Describe how each algorithm works. How did you go about implementing them? Also provide details on how and why you selected the hyperparameters you did (justify any choices here empirically).
   For example, if you used regularised linear regression:
   - Why did you choose this algorithm, and why did you add regularisation?
   - What value of λ did you use? Why was this a good choice (with evidence)?
   - What basis functions did you use? Why?
   - How did you train the model? e.g. gradient descent with $\alpha = 0.2$. Why did you choose this?

(4) Present your results. At the very least, format your errors in the form of a confusion matrix. You can talk about other performance metrics as well. Analyse these results: what kind of problems may there be, and how could you deal with them? Also compare your algorithms to a random baseline: how well would you do if you were randomly guessing the class?

(5) Very important: a discussion and comparison of your results from the various algorithms. E.g., what worked best/worst and why you think this is so. What is the best possible performance you can achieve on this dataset? What kinds of data points did each algorithm struggle with? How did you do that? What would you recommend someone else try if they were interested in working with this data? Make sure all your points here are substantiated with evidence from your experiments. Also include a discussion of the times taken to run each algorithm.

(6) Upload your code as well. This code will be plagiarism checked! Note: you may use existing libraries for tasks such as file handling and data processing, but the machine learning algorithms must be coded yourself!

Notes:

- Your dataset must be sufficiently large and with enough attributes. Credit will be given for an interesting dataset.
- **You must use your own implementations of the core algorithms, but can use helper libraries for other functions**. Be explicit about what you used, and cite where appropriate.
- The more algorithms you try, the better.
- If you have some nice visualisations/graphs, please include them.
- Although you must submit your code, **you will only be marked based on what is in your ONE pdf file**. Anything not in here will not get you marks, and if you do not submit a pdf, you will get zero.
- Make sure your document is well formatted, and the images are clearly visible.
- Don't just paste code in, but describe what you have done in each case.
- Include any references you have used at the end of the document.

Important:

- You need to discuss your dataset with me by **Friday, 14 May**.
- The closing date for submission is the end of **Thursday, 17 June**.
- You must submit and work in groups of **between two and four people**. Make sure all your names AND student numbers are on the submission, otherwise you will receive 0.