COMS 3007: Machine Learning Assignment 2020

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, but any source is suitable: https://archive.ics.uci.edu/ml/datasets.html
This MUST be a classification problem.
Apply AT LEAST TWO supervised learning algorithms to your dataset. Bonus marks will be given for more.

You must **submit a PDF document as well as your code** to Moodle containing the following information:

(1) A description of your dataset: what are the attributes, what are the targets, how many datapoints do you have, and some sample datapoints from the dataset. State what you are trying to predict with the data.

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, and the split into training/validation/test data.

(3) A list of classification algorithms you applied to the data, together with the details of each implementation and the error on the test set. Also provide details on how and why you selected the hyperparameters you did. Present the errors at least in the form of a confusion matrix.

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) A brief discussion 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? How did you do that? What would you recommend someone else try if they were interested in working with this data?

(5) 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.

Important:

- You need to discuss your dataset with me by **Friday, 3 April**.
- The closing date for submission is the end of **Friday, 15 May**.
- 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.