EE3-23: Coursework for Introduction to Machine Learning

Spring, 2018

1 Description

The goal of the coursework is to assess your ability to solve a machine learning project. There are multiple project options, you have to complete one of them. In each project, you are given a single dataset. Each instance consists of features and a variable to be predicted.

You are permitted to use any programming language (recommendation: matlab or python), including any standard libraries available. However, you are not allowed to use custom code available on the internet developed to solve the dataset of your choice.

Your task is to obtain a predictor with as small test error as you can. However, while the performance of your final predictor will be taken into account, the main component of the assessment is the process you used to obtain your solution. As such, you are required to perform the following steps:

(a) **[10]** Present the problem of your choice in a formalised way, choose a loss function that reflects the potential use of the predictor (see lectures).

(b) **[20]** Propose and implement baseline predictors/classifiers and methods to train them, e.g., include a linear method from the course. Present your findings.

(c) **[20]** Propose more advanced algorithms to solve the problem.

(d) **[20]** Implement the methods proposed in (c), give insights into the training and evaluation process.

(e) **[20]** Asses their performance and present your proposed solution to the problem.

(f) **[10]** Discuss your overall findings and conclusions.

You will need to submit a written report and the code used to generate the results (see below). The coursework will be assessed based on the report; the code might be used to check the validity of the report and the originality of the solution, but it will not be assessed.

2 Datasets

Choose one of the following datasets. These datasets may need some minor curating, which should be discussed in the report if done. In particular, create a reasonable train-test split for your chosen dataset.

(a) https://archive.ics.uci.edu/ml/datasets/spambase
Here you need to predict if an email is spam or not.

(b) https://archive.ics.uci.edu/ml/datasets/Wine+Quality
Here you need to predict the wine quality.

(c) https://archive.ics.uci.edu/ml/datasets/SMS+Spam+Collection
Here you need to predict if an SMS is spam or not.

(d) http://archive.ics.uci.edu/ml/datasets/Beijing+PM2.5+Data
Here your task is to estimate PM2.5 concentration at 8:00am on any day based on data available at 8:00am at the previous day.

(e) You can propose your favorite dataset and problem, but you need to receive an approval to ensure the problem is of adequate quality.

3 Report

The report should be no longer than **4 A4 pages with font size 10pt** (use the IEEE standard double column paper format, either in MS word or latex in Blackboard). List of references and appendix do not count for this page limit.

General principles for writing technical report are expected to be known and adhered to. Similarly for practices in conducting experiments, some are as listed below:

- The choice of critical components in your approach should be clearly justified.
- The approach and experiments should be described such that there is no ambiguity in the settings, protocol and metrics used.
- Select relevant results that support the points you want to make.
- The important results should be in the report, not just in the appendix. Any results which are used in your decision process should be reported.
- Use clear and tidy presentation style, consistent across the report, e.g., figures, tables.
- The main points are made clear, identifying the best and the worst case results or other important observations.
- Do not copy standard formulas from lecture notes, explain algorithms taught in the class in detail, or copy figures from other sources. References to lecture slides or publications/webpages are enough in such cases, however short explanations of new terms or parameters referred to are needed, as well as the derivations of new formulas (if any).

Include the following pledge into your report:
I, <YOUR NAME>, pledge that this assignment is completely my own work, and that I did not take, borrow or steal work from any other person, and that I did not allow any other person to use, have, borrow or steal portions of my work. I understand that if I violate this honesty pledge, I am subject to disciplinary action pursuant to the appropriate sections of Imperial College London.
If you don't include your pledge, you will get 0%.

4 Submission

Submit a single zip file named `<username>.zip` e.g. `km2316.zip`, in Blackboard. The file should contain :

(a) report in PDF (use template from Blackboard), name it `<username>.pdf`.

(b) code and other relevant files in directory `./code`.

If your code is in matlab or python, name it `<username>.m` or `<username>.py`, otherwise provide a make file which produces an executable `<username>` in a linux system. Running your code should generate all the plots required in the solution. Data files are assumed to be in the folder `../data/` relative to the source code but do not include the actual data files in your submission. Instead include a short text file describing the content of `./data/` directory.

5 Deadline: 23 March 2018, 6:00pm