COMS30069: Machine learning coursework 2022

1 Introduction

This coursework is designed for you to apply some of the methods that you have learned during our Machine Learning unit and that are also commonly applied in practice. Given that this is your only assessment for this unit the coursework is designed to be relatively open-ended with some guidelines, so that you can demonstrate your knowledge of what was taught – both in the labs and in the lectures.

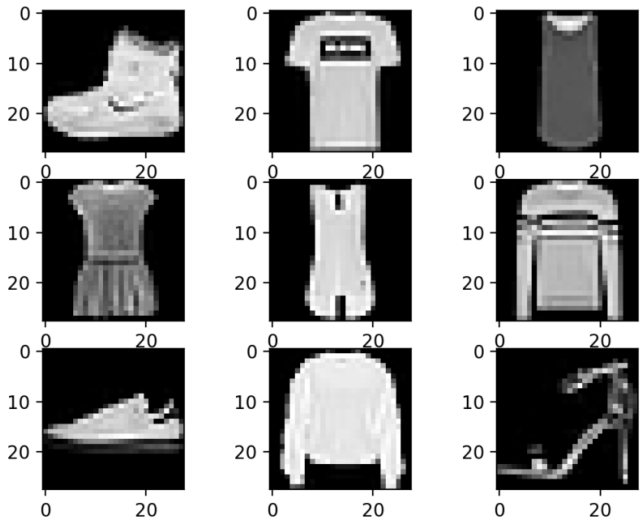

Figure 1: Samples from the fashion MNIST dataset.

2 Tasks

In this coursework, we will focus on modern yet accessible machine learning dataset: fashion MNIST dataset[1] and the California housing regression dataset[2]. We recommend that you first get a basic implementation, and start writing your report with some plots with results across all four topics, and then gradually improve them. Where suitable you should discuss your results in light of the concepts covered in the lectures (e.g. curse of dimensionality, overfitting, etc.). Feel free to use a subset of the full fashion MNIST dataset to speed up optimisation. You should use this coursework to demonstrate your understanding of the different machine learning models and concepts not to get the best performance. For a **good grade you must introduce and explain all the methods used using equations and/or algorithms as appropriate**.

2.1 Analysing fashion-MNIST (25 marks)

To gain a deeper understanding of a particular dataset it is often a good strategy to analyse it using unsupervised methods. Use **only** the fashion-MNIST dataset for this task.

2.1.1 PCA (10 marks)

Run PCA on the fashion-MNIST dataset.

1. How much variance do the first and second principal components explain? (5 marks)

2. Create a 2D scatterplot of the data projected onto the first two principal components. Differentiate between the different classes using colour. To what extent do the datapoints in your scatter plot cluster into the different classes? (5 marks)

2.1.2 Gaussian Mixture Modelling (15 marks)

1. Use a Gaussian Mixture Model (GMM) to do (soft) clustering using just the first two components from the PCA analysis above. Make the

[1] https://github.com/zalandoresearch/fashion-mnist

[2] https://scikit-learn.org/stable/datasets/index.html#california-housing-dataset

GMM a mixture of 10 Gaussians. Plot your clusters in 2D. It is up to you to choose how best to plot clusters. (8 marks)

2. Analyse the relationship between the clusters produced using the Gaussian Mixture Model and the class labels. It is up to you to decide what should be included in this analysis. (7 marks)

2.2 Classifiers (25 marks)

Building on what you learnt from the labs, here you are asked to contrast two types of classifiers, Artificial neural networks (ANNs) and Support Vector Machines (SVMs). Using the libraries used during the labs you only need to run two classifiers, and discuss its advantages and disadvantages over the other. You should make sure to control for overfitting. Use **only** the fashion MNIST dataset for this task. Feel free to use a subset of the full dataset to speed up optimisation.

2.2.1 Artificial neural networks (10 marks)

Here you are going to study and discuss ANNs as a model for the fashion MNIST classification dataset. In particular you should:

1. Train an ANN, plot the training and validation learning curves. Do you see any signs of overfitting? Interpret and discuss your results. (1 mark)

2. What are your results in the testing dataset? Interpret and discuss your results. (2 marks)

3. How sensitive is this method to different hyperparameters? Make use of plots to help you discuss this point. (5 marks)

4. Plot decision boundaries and discuss their relevance. (2 marks)

2.2.2 Support Vector Machines (15 marks)

Here you are going to study and discuss SVMs as a model for the fashion MNIST classification dataset. In particular you should:

1. Train an SVM (with a specific Kernel), plot the training and validation learning curves. You may need to subsample the dataset if SVM training is taking too long. Do you see any signs of overfitting? Interpret and discuss your results. (1 mark)

2. What are your results in the testing dataset? Interpret and discuss your results. (2 marks)

3. How sensitive is this method to different hyperparameters? For example the different types of kernel (e.g. linear, RBF, etc.). Make use of plots (e.g. performance on test dataset as a function of different hyperparameters) to help you discuss this point. (5 marks)

4. Plot decision boundaries and discuss their relevance. (2 marks)

5. Compare your SVM results with the ANN above in terms of performance and the time it takes to train each method. For example, use bar plots to compare their performances and training times next to each other. Which is the better model? And why? (5 marks)

2.3 Bayesian linear regression with PyMC (25 marks)

In this task you are required to use PyMC to perform Bayesian linear regression on the California housing dataset which is easily available via the sklearn.datasets.fetch_california_housing function. The goal with this dataset is to predict the median house value in a 'block' in California. A block is a small geographical area with a population of between 600 and 3000 people. Each datapoint in this dataset corresponds to a block. Consult the scikit-learn documentation for details of the predictor variables.

1. Produce a suitable plot which shows how longitude and latitude affects median house price. Explain what your plot tells you about the relationship between these two predictors and median house price. (5 marks)

2. Decide whether you need to transform and/or clean the data. Justify your decision in your report. (5 marks)

3. Choose prior distributions for all model parameters and then use PyMC to get approximate posterior distributions over each model parameter.

In your report include plots of these posterior distributions, as well as giving the mean and standard deviation of each distribution. (5 marks)

4. Did your run of PyMC succeed, that is: did it produce good approximations to the desired posterior distributions? Explain your answer. (5 marks)

5. The full dataset has 20640 datapoints. Now run PyMC on two random samples of datapoints, one of size 50 and one of size 500. Compare the posterior distributions you get for the 3 dataset sizes (50, 500 and 20640), stating what the most important differences are and explaining how these differences arose. (5 marks)

2.4 Trees and ensembles (25 marks)

This part extends the work on decision trees and ensemble methods from lab 7 to regression on the California Housing regression task.

2.4.1 CART Decision Trees (15 marks)

In this part you are going to apply a decision tree regressor to the California housing dataset and analyse its behaviour. Your answers should address the following points:

1. Briefly explain how the CART decision tree method works. (2 marks)

2. Use model selection to optimise the hyperparameters of the model. Which hyperparameter has the strongest effect on the model's performance? Use a plot to show this effect. (5 marks)

3. How do the hyperparameters affect the training time? Use plots to support your discussion. Explain how this affects your choice of hyperparameter values. (3 marks)

4. What are the results for your chosen setup on the test set? Interpret and discuss your results. (3 marks)

5. In what situations could a decision tree be a better choice than linear regression? (1 marks)

6. Identify a data point that the model classifies incorrectly. Explain how the model classifies this data point, in terms of the sequence of decisions it makes. Use a diagram to help with your interpretation. (2 marks)

2.4.2 Ensemble Methods (10 marks)

Here, you need to choose an ensemble method and apply it to the California housing dataset. Your answers should address the following points:

1. Briefly explain how your chosen ensemble method works, mentioning the key benefits of your chosen method. You do not need to give an exhaustive set of equations. (2 marks)

2. How is your method affected by the number of base models in the ensemble? Use a plot to support your discussion. (3 marks)

3. What are the results for your chosen setup on the test set? Interpret and discuss your results, and briefly compare them with those of the single decision tree and Bayesian linear regression. Use plots or tables to support your comparison. (5 marks)

3 Implementation

You are expected to build on the skills you have learned during the labs. Therefore, you should use the Python libraries used during the labs, namely Scikit-learn and PyMC. You can use other libraries, but we won't be able to provide support on those.

4 Assessment criteria

Your coursework will be evaluated based on a submitted report, containing the appropriate discussion and results. The aim of this report is to demonstrate your understanding of the methods you used and the results that you have obtained. Note: In the report it is important that you briefly describe the methods used.

The report should be no more than **10 pages long**, using no less than **11 point font** (excluding references). Note that your report should be quality rather than quantity, so do not feel like you have to use 10 pages if they are

not needed. If you wish to use a template for Latex, you can use the basic report template or the Coling 2020 template. Submission: On Blackboard (under Assessment, Coursework) with a **pdf file (as cw_userid.pdf) for the report together with your code (e.g. with the Jupyter Notebooks you have used; as cw_userid.zip)**. Note that your code is not going to be used for marking, only to validate your work.

To gain high marks your report will need to demonstrate clearly a thorough understanding of the tasks and the methods used, backed up by a clear explanation (including figures) of your results and analysis. The structure of the report and what is included in it is your decision and you should aim to write it in a professional and objective manner so that it addresses the issues mentioned above. In particular you need to explain clearly the following elements:

1. Analyse the fashion MNIST dataset using GMMs **and** PCA (25%)
2. Apply and discuss the results of a classifier on fashion MNIST (ANN **and** SVM) (25%)
3. Bayesian linear regression on the California housing dataset (25%)
4. Implement random forest and stacking and contrast (using the California housing dataset) them with the previous methods in terms of performance and interpretability (25%)

Deadline: The deadline for submission is **13:00 (1pm) on Thursday 8th December 2022 (end of week 11)**. Students should submit all required materials to the "Assessment, submission and feedback" section of Blackboard - it is essential that this is done on the Blackboard page related to the "With Coursework" variant of the unit.

5 Support provided

This is your only form of assessment so we cannot provide direct support on the coursework. However, we can clarify questions you might have about specific material from the lectures and/or the labs. We will try our best to be available for this via Teams on *Tuedays 1-2pm* and at least one of the lectures will be available *9-10am on Thursdays*, in the usual lab space.

6 Further clarifications

- You are expected to work around 8h/day for 5 days/week.
- Feel free to use the labs materials as a starting point.
- To make the best use of space you should use matplotlib subplots and use a given plot to make comparisons (e.g. training and validation learning curve).
- We suggests that you use Python with Jupyter Notebook and the libraries that we used during the labs (e.g. Scikit-learn and PyMC). You can use others but not that support wont be available.
- **Academic offences**: Academic offences (including submission of work that is not your own, falsification of data/evidence or the use of materials without appropriate referencing) are all taken very seriously by the University. Suspected offences will be dealt with in accordance with the University's policies and procedures. If an academic offence is suspected in your work, you will be asked to attend an interview with senior members of the school, where you will be given the opportunity to defend your work. The plagiarism panel are able to apply a range of penalties, depending on the severity of the offence. These include: requirement to resubmit work, capping of grades and the award of no mark for an element of assessment.
- **Extenuating circumstances**: If the completion of your assignment has been significantly disrupted by serious health conditions, personal problems, periods of quarantine, or other similar issues, you may be able to apply for consideration of extenuating circumstances (in accordance with the normal university policy and processes). Students should apply for consideration of extenuating circumstances as soon as possible when the problem occurs. If your application for extenuating circumstances is successful, it is most likely that you will be required to retake the assessment of the unit at the next available opportunity.

7 Marking guidelines

7.1 Outstanding (80+)

+ mastery of advanced methods in all aspects;
+ truly impressive outcome, novelty, with strong research elements – close to publication quality;
+ synthesis in an original way using ideas from the unit but also from the literature;
+ outstanding presentation of work, with very clear description of the methods and results;
+ excellent use of plots to support the interpretations;
+ evidence of outstanding unique and individual contributions.

7.2 First class (70+)

+ excellent outcome in all aspects;
+ evidence of excellent use and deep understanding of a wide range of techniques;
+ study, originality and synthesis clearly beyond the minimum requirements set out in the coursework description;
+ excellent presentation of work, with very clear description of the methods and results;
+ very good use of plots to support the interpretations;
+ evidence of excellent contributions or insights into the methods tested.

7.3 Merit (60+)

+ very good outcome with complete solutions for all the required aspects of the assignment;

+ evidence of very good use and strong understanding of a range of techniques;

+ study, comprehension and synthesis fully meet or exceed the requirements set out in the coursework description;

+ very good presentation of work, with clear description of the methods and results;

+ good use of plots to support the interpretations;

+ evidence of critical analysis and judgement of the methods tested.

7.4 Good (50+)

+ good outcome but some of parts of the assignment not fully completed;

+ evidence of good use and understanding of standard techniques;

+ some grasp of issues and concepts underlying the techniques;

+ adequate presentation of work, including a description of the methods and results;

+ some good use of plots to support the interpretations but with some notable shortcomings;

+ evidence of understanding and appropriate use of techniques.

7.5 Passing (40+)

+ Limit outcome yet basic, partly solutions to all the 4 main topics

+ limit understanding as demonstrated through discussion and plots

+ poor presentation of results