COMS30034: Machine learning coursework

December, 2020

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 MNIST dataset.

2 Tasks

In this coursework, we will focus on the classical hand-written 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.).

2.1 Analysing MNIST

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

2.1.1 PCA

Run PCA on the MNIST dataset. How much variance does each principal component explain? Plot the two components that explain the most variance. Interpret and discuss your results.

2.1.2 K-Means

Apply K-means (with $K = 10$) using the first two components from the PCA analysis above. Plot your clusters in 2D and relate them to the digit classes. What does each cluster correspond to? How good is the match between a given cluster and a specific digit? Interpret and discuss your results.

2.2 Classifiers

Building on what you learnt from the labs, here you are asked to contrast two types of classifiers, ANNs and 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 MNIST dataset for this task.

[1] http://yann.lecun.com/exdb/mnist/

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

2.2.1 ANNs

Train an ANN, plot the training and validation learning curves. Does the model overfit? What are your results in the testing dataset? Interpret and discuss your results. How do they compare with SVMs? How do the hyperpameters (e.g. learning rate) impact on performance?

2.2.2 SVMs

Train an SVM (with a chosen Kernel) and perform the same analyses as for ANNs. You may need to subsample the dataset if SVM training is taking too long. Interpret and discuss your results. Does the model overfit? How do they compare with ANNs? And why? How does the type of kernel (e.g. linear, RBF, etc.) impact on performance?

2.3 Bayesian linear regression with PyMC3

In this task you are required to use PyMC3 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.

As always with Bayesian analysis it is up to you to choose your prior distributions. Be sure to justify your choice of priors in your report. What do the results produced by PyMC3 tell you about what influences house value in California? Is it necessary and/or useful to transform the data in some way before running MCMC?

2.4 Building an ensemble

Here, you will implement both of the following steps for the California Housing regression task.

2.4.1 Random Forest

This part builds on the related lab (week 7). First, run a random forest regressor for the California housing dataset, and contrast this with your

previous Bayesian linear regression method. For this you can use the RandomForestRegressor class from Scikit-learn.

Analyse the effect of the hyperparameters of the random forest, such as the number of estimators (or *base models*, i.e., the number of decision trees that are combined into the random forest). Look at the constructor of the RandomForestRegressor class to see what hyperparameters you can set. In your analysis, include the following plots and discussions but you may wish to add further analysis of your own:

1. Plot the relationship between a hyperparameter and the performance of the model.
2. Optimise the hyperparameter on a validation set.
3. Plot the trade-off between time taken for training and prediction performance.
4. What do you think is a good choice for the number of estimators on this dataset?
5. What is the effect of setting the maximum tree depth or maximum number of features?
6. Is the random forest interpretable? Are the decision trees that make up the forest interpretable?

2.4.2 Stacking

Bayesian linear regression and decision trees are two very different approaches to regression. Ensemble methods can exploit such diversity between different methods to improve performance. So now you will try combining the random forest and Bayesian linear regression using *stacking*. Scikit-learn includes the StackingRegressor class to help you with this. In the report, explain the stacking approach and describe your results, making sure to cover the following points:

1. When does stacking improve performance over the individual models (e.g., try stacking with a random forest with max_depth = 10 and n_estimators = 10)?

2. What happens if we just take the mean prediction from our base models instead?

3. Use a DecisionTreeRegressor as the final_estimator and visualise the tree to understand what stacking is doing.

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 PyMC3. 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**. 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 MNIST dataset using K-means **and** PCA (25%)
2. Apply and discuss the results of a classifier on 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%)

Suggestion for discussion points: after describing each method you use, consider what you would expect the results to look like, e.g., how you think K-means clusters will relate to the digits data. Then use your results to validate your hypothesis and look for patterns of errors that each method makes. Can you explain why each method makes certain types of error?

Deadline: The deadline for submission of all optional unit assignments is 13:00 on Friday 11th of December. 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.

Working in pairs: You are encouraged to work in pairs or small groups, but both report and code need to be your own pieces of work.

5 Support provided

TAs and Lecturers will be available to provide support during the usual lab hours (Thursdays from 10am to 1pm as in your timetable). This will run as drop-in clinic where you can just join one of the booths [blue, green, orange, ...]. There is going to be a TA per room. Note that these sessions are there to help with conceptual understanding rather than helping you fix bugs. Also, we cannot provide advice about whether what's in your report is enough or not.

6 Further clarifications

- 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).
- You should use Python with Jupyter Notebook and the libraries that we used during the labs (e.g. Scikit-learn and pyMC3)
- **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, using the following online form. You should note however that extensions are not possible for optional unit assignments. 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 project (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 project (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 project (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 project (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 project (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