COMS30077: Machine learning coursework 2023

1 Instructions

Deadline: Thursday 7 December. See Section 7.1 for further details on the deadline.

1. Your submission should consist of a single PDF which is your report and one or more Jupyter notebook files.
2. Where a task requires you to write Python code to complete, be sure to include that code in a Jupyter notebook file (but there is no requirement to include the code in your report).
3. You can and should import appropriate libraries (e.g. scikit-learn, PyMC) to help you complete the tasks you are required to do.
4. Your Jupyter notebook file(s) should run without throwing errors.
5. Your Jupyter notebook file(s) do not need to include output cells (but it's fine if they do).
6. For your submission you are asked to produce various plots. These plots should be included in your report and the code required to produce them should be included in your Jupyter notebook file(s).
7. Feel free to use the lab materials as a starting point.
8. To make the best use of space you should use matplotlib subplots or show multiple lines in a single plot when making comparisons.
9. We suggest using the libraries that we used during the labs (e.g. Scikit-learn and PyMC). You may use others but we will not be able to provide support for them.

10. Your PDF report should be be no more than **8 pages long** (excluding references) and no less than 11 point font. You can find suitable templates here.

11. Name your report file *cw_⟨userid⟩.pdf* and package it into a zip file with all your code, named *cw_⟨userid⟩.zip* (replace *⟨userid⟩* with your university username).

2 Classification and Sequence Labelling (50 marks)

In this coursework you will apply some of the methods learned during the unit, and explore their behaviour using some real-world datasets. For the first set of tasks, we will be working with the dataset "Activity recognition with healthy older people using a batteryless wearable sensor". This dataset contains sequences of readings taken from several sensors on a device worn by a person as they move about in a room. The goal is to predict the activity label for each data point, i.e., what the person was doing, which may be 1: sit on bed, 2: sit on chair, 3: lying, or 4: ambulating. To help you access the data, we provide a notebook "activity_recognition.ipynb" in our Github repository (the same repository as the lab worksheets). The notebook code loads the data into a suitable format for the following tasks.

Task 1 Implement and evaluate a neural network classifier for the activity recognition dataset. For this task, ignore the sequence of data points and use the *classification* data provided in the notebook. Your report should cover the following: (12 marks total)

(a) Train a neural network classifier and plot training and validation learning curves. Interpret and discuss your plots, considering over/underfitting. (3 marks)

(b) Present and briefly discuss your results on the test set. (2 marks)

(c) Which hyperparameters have the strongest effect on model performance? Use plots to help you investigate this. (7 marks)

Task 2 Implement and evaluate an ensemble method using decision trees as base models. Again, use the activity recognition *classification* data

provided in the notebook and ignore the sequence of data points. Your report should cover the following: (18 marks total)

(a) Train an ensemble of decision trees, and briefly describe how your ensemble improves performance over a single model. (2 marks)

(b) Present and briefly discuss your results on the test set. (2 marks)

(c) Show the relationship between the error rate of the ensemble, the error rate of individual base models, and the number of base models. Does ensembling improve performance? (7 marks)

(d) How sensitive is the model to different hyperparameters (e.g., maximum tree depth)? Use plots to explore the different trade-offs. (7 marks)

Task 3 The previous tasks ignored the ordering of data points. Here, we will use the *sequence labelling* data in the notebook to label each sequence of observations, taking their order into account. Your task is to implement and evaluate a sequence labelling method. (20 marks total)

(a) Train a sequence labeller, present your results on the test set and briefly explain the results. (3 marks)

(b) Plot a transition matrix and explain what the matrix tells us about the relationships between different activities. (6 marks)

(c) Can you identify any features that are particularly informative when predicting activity labels? Justify your answer, e.g., by comparing the means of the emission distributions. (4 marks)

(d) Compare your sequence labeller with the neural network and decision tree ensemble, considering performance, training time, and interpretability. What are the strengths and weaknesses of each approach? You may use both plots and discussion in your comparison. (7 marks)

3 Clustering and dimensionality reduction (15 marks)

Download the Breast Cancer Wisconsin (Diagnostic) dataset from here. Do **not** use the "IMPORT IN PYTHON" option, just grab the zip file and unzip

it. You will end up with two files: `wdbc.data` which is the data itself and `wdbc.names` which describes the data. As you can read, the data can be used to learn a classifier which uses 30 real-valued features to distinguish between benign and malignant tumours.

In this section you are asked to compare the clusters produced by Gaussian mixture clustering to the true classes. But instead of clustering the data as is, you are first asked to use PCA to reduce the dimensionality down to 2 dimensions.

Task 4 Use PCA to reduce the data (without the class labels) from 30 dimensions to 2 dimensions and produce a scatter plot of this 2-dimensional data using colours to indicate the class label of each data point. In addition, in your report, state how much variance is explained by the first principal component and how much is explained by the second. (5 marks)

Task 5 Fit a Gaussian mixture model with 2 components using the 2-dimensional data and display the (soft) clusters using a scatter plot. The colour of each point in the scatter plot should represent the *responsibilities* associated with that point—so there will be more than 2 colours in your scatter plot. (The rightmost plot on slide 6 of the lecture on k-means and mixtures of Gaussians is an example of this sort of plot.) (5 marks)

Task 6 You now have 2 different sorts of (coloured) scatter plot. Identify the main differences between them and explain these differences. (5 marks)

4 Classification with SVMs (10 marks)

Task 7 Use an SVM approach to learn a classifier from the first 450 datapoints of the Breast Cancer Wisconsin (Diagnostic) dataset and report its accuracy on the the last 118 datapoints (i.e. use these 118 datapoints as a test set). Attempt to maximise the test set accuracy by choosing an appropriate kernel and appropriate degree of regularisation. Feel free to score candidate SVM predictors on the test set to see what works best (so the test set is really acting as a validation set, not a test set in the strict sense). (4 marks)

Task 8 Repeat the previous task but now use the 2-dimensional dataset (plus class labels) produced earlier via PCA. (4 marks)

Task 9 Discuss to what extent, if any, using the 2-dimensional approximation to the original data affected test set accuracy. (2 marks)

5 Bayesian Linear Regression (25 marks)

Download the Seoul Bike Sharing Demand dataset from here. You are required to use Bayesian Linear Regression (BLR) where the variable to be predicted is the number of bikes rented. Although the posterior distributions produced by BLR could be used to make predictions, you are not required to do that for this coursework; instead you are asked to produce these posterior distributions and analyse them.

The first line of the downloaded data gives the names of the columns. Two of these names are temperatures which contain the ° symbol. This symbol may cause you problems when reading in the data, so just edit the file to delete it.

Task 10 It is clearly not possible to do linear regression without altering the downloaded data in some way (for example, a number of variables are not numeric). Bearing in mind that the goal is to predict number of bikes rented, create a new dataset from the original downloaded dataset that is suitable for Bayesian Linear Regression (BLR). In your report, describe and justify how you have transformed the data. (8 marks)

Task 11 Choose prior distributions for all parameters in your BLR model and justify your choices (in your report). (2 marks)

Task 12 Use PyMC to perform Bayesian Linear Regression on your new dataset. Since you may well be using many predictor variables it would be tedious to construct your BLR model using the approach we used in Section 3.2 of Lab 3. Instead you will find it more convenient to use the Model Specification approach used here which uses the `coords` keyword. (You can use the approach we used in Lab 3 if you want, with no loss of marks, but the Lab 3 method is more hard work.) Use `arviz.summary` to generate summary statistics for the (approximate) posterior distributions for each parameter in your model. Include this

summary in your report. (You do not need to include plots such as those produced by `arviz.plot_trace` in your report.) (5 marks)

Task 13 Comment, in your report, on whether you have evidence that the MCMC sampling has generated reasonable approximations to the posterior distributions of your parameters. (2 marks)

Task 14 The summary generated by `arviz.summary` gives you a posterior mean value for each model parameter. What do these values tell us about what influences the number of bikes hired? Are any of these values surprising? (e.g. are there any examples of negative values associated with variables you would expect to have a positive effect on the number of bikes rented?). (5 marks)

Task 15 Discuss whether linear regression is a suitable model for this type of data. (3 marks)

6 Support provided

This is an 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. There are Coursework Support Sessions in MVB 2.11 1000-1200 on the following Thursdays: 23 Nov, 30 Nov, 7 Dec. You can also ask questions on Teams using the following Teams group: "COMS30077: Machine Learning (with Coursework) 2023/24 (TB-1, A)".

7 General instructions for coursework

7.1 Deadline and submission process

The deadline is midday on Thursday 7th of December. Blackboard will start applying late penalties at 1pm sharp. You should aim to submit by midday, and treat the extra hour as a buffer for serious emergencies, not a target. If you aim for midday and something goes seriously wrong you can then still contact the school office well before 1pm. You should not attempt to submit by 'almost' 1pm because if you go even slightly over then you will get 10 marks deducted (so a 65 would become a 55).

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.

7.2 Time commitment

You are expected to work on both of your optional unit courseworks in the 3-week coursework period as if it were a working week in a regular job—that is 5 days a week for no more than 8 hours a day. The effort spent on the assignment for each unit should be approximately equal, being roughly equivalent to 1.5 working weeks each. It is up to you how you distribute your time and workload between the two units within those constraints.

You are strongly advised NOT to try and work excessive hours during the coursework period: this is more likely to make your health worse than to make your marks better. If you need further pastoral/mental health support, please talk to your personal tutor, a senior tutor, or the university wellbeing service.

7.3 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. Further information on the university's academic integrity policy can be found below: https://www.bristol.ac.uk/students/support/academic-advice/academic-integrity/

7.4 Extensions

If you are unwell, or there is another reason why you are unable to meet a due date, you can request an extension, however you should plan your work so

that your submission is not delayed by short-term circumstances such as minor illness. Further information and guidance on how to request an extension can be found on the below link: https://www.bristol.ac.uk/students/support/academic-advice/assessment-support/request-a-coursework-extension As of the 23/24 academic year the deadline for the submission of an extension request is 48 hours before the coursework submission deadline. If the extension deadline has passed, then please refer to the guidance on exceptional circumstances.

7.5 Exceptional 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 exceptional circumstances (in accordance with the normal university policy and processes). https://www.bristol.ac.uk/students/support/academic-advice/assessment-support/exceptional-circumstances/#:~:text=Exceptional%20circumstances%20are%20unexpected%2C%20unavoidable,academic%20performance%20in%20an%20assessment Students should apply for consideration of exceptional circumstances as soon as possible when the problem occurs, using the following online form: https://www.bristol.ac.uk/request-extenuating-circumstances-form 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.

8 Assessment Criteria

Your coursework will be evaluated based on the report, with the code required to provide evidence of your work. To gain high marks, your report should demonstrate a thorough understanding of the tasks and methods used, backed up by clear explanations, including figures, of your results and analysis. The report should be quality, rather than quantity, so you do not have to fill the report up to the page limit. Your report should introduce and briefly all methods used, with equations where appropriate. Where suitable you should discuss your results in light of the concepts covered in the lectures (e.g. curse of dimensionality, overfitting, etc.).

9 Marking guidelines

9.1 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.

9.2 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.

9.3 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.

9.4 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.

9.5 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