Data Science Toolbox Portfolio Questions

Long form questions

Daniel Lawson — University of Bristol

due Week 12

Part of Data Science Toolbox.

Deadline: Week 12 (See Timetable for details).

Data Science Portfolio Overview

- Worth 40% of the course
- Individually assessed, follow UoB coursework assessment guidelines.

What is a portfolio?

- A portfolio in Data Science Toolbox part of your ongoing assessment.
- You should attempt to work through it in time with the course material, then come back and finesse your work at the end of the Unit.
- It is composed of one section per Block.
- Each section consists of two components:
  - **Worksheets**: Multiple choice questions, submitted via **Noteable**.
    - You must answer **every** worksheet.
  - **Longform**: Deeper descriptions of material you have examined during the Worksheets and Workshop.
    - You must answer **5** of the available **10** portfolios.
- The Longform content should be no more than 1 page per block.

Portfolio Content

Longform content is linked **within each Block** but is also available in the Assessments page. Worksheets are found within Noteable.

Guidance on Individual Portfolios

The Portfolio is assessed for blocks 2-11. Block 1 is marked similarly but is formative, i.e. does not contribute to your mark. The deadline is in assessment preparation week of TB1. In each block contains two activities:

1. Multiple choice questions submitted via Noteable (log in via Blackboard). These should be straightforward, either direct from your notes or with simple experiments you can conduct as extensions of the Workshop. These are worth 20% of the Portfolio mark.
2. Long-form reflective questions that should require a deeper understanding of the course material and may require you to undertake further reading or experimentation. These are worth 80% of the Portfolio mark.

You may take the multiple-choice component at any time and it is recommended that you do this when you work through the Workshop content. The long-form content is submitted at the end of the course, and you are recommended to make a first draft/note form attempt when you first see the content, and reflect back on it in a finessing stage during the examination preparation time (in lieu of an exam).

Length and format of long-form portfolio

Your Portfolio should give a **one-page** answer to one question of your choice from 5 Blocks (2-11). Therefore the whole Portfolio is only 5 pages long. However:

- The goal is not to make you undertake a length-finessing exercise. If the content you provide appears as if it would fit on one page after such an exercise, you can submit is anyway. **There is a strict limit of 8 pages** for the portfolio content, with answers that are clearly too long being penalised.
- You can however submit **Supporting Evidence** as an appendix to the portfolio. It will not be directly assessed but may be used as evidence to support your claims, i.e. any statements you make with supporting evidence will be more favourably interpreted, but if your statements are carefully given and correct the evidence is not essential. This is not limited. Appropriate content is RMarkdown files knitted to pdf, Jupyter Notebooks, etc.

Portfolio Questions:

Questions are released on a rolling basis. Check back on this document for additional content as it is released, or see Assessments for individual documents.

Data Science Toolbox Portfolio Questions

02 Regression and Statistical Testing

Daniel Lawson — University of Bristol

Block 2

Portfolio 02

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R02.1: It was claimed without proof in Lecture 02.1 that the leave-one-out cross validation error can be cheaply computed for linear regression as:

$$CV = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{e_i}{1-h_{ii}}\right]^2,$$

where $e_i = y_i - \hat{y}_i$, $\hat{y}_i = \beta X_i$ and h_{ii} is the diagonal entries of the hat matrix. This also works for penalised regression, to come later. Consider the proof presented in https://robjhyndman.com/hyndsight/loocv-linear-models/ or otherwise, and rewrite this proof with simple annotations for an Undergraduate audience. Briefly discuss the implications of the theorem for both the Temperature and Diamonds datasets from Workshop 2.3.

Question R02.2: Consider the paper Model-agnostic out-of-distribution detection using combined statistical tests. What are the key reasons that justify statistical testing over standard machine-learning, and how does it relate to our course content? Provide an experimental design to test whether these results hold in practice for the experiments performed in Diamonds data from Workshop 2.3.

Question R02.3: Imagine that you are tasked with making a temperature prediction for 2040 based on the Temperature Data used in Workshop 2.3. Design and justify a cross-validation setup that could be used to obtain predictions along with uncertainty quantification, carefully describing its advantages over what is presented above, and its limitations. You may wish to investigate standard forecasting methods.

Data Science Toolbox Portfolio Questions

03 Latent Structures, PCA, and Clustering

Daniel Lawson — University of Bristol

Block 3

Portfolio 03

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R03.1: Read the (python) documentation for Sparse SVD. How is this making computational efficiencies over a standard SVD and when might this fail? The documentation discusses the sense in which the results are accurate in terms of "subspace_angles". Discuss what this means, and how it relates to the use of the SVD in PCA.

Question R03.2: Read the documentation about how HDBSCAN works. Reflect on the importance of dimension in this for the construction of the nearest-neighbour step. You might want to refer to results in the literature such as "When is Nearest Neighbor meaningful?".

Question R03.3: Imagine that you are trying to understand the Cyber Security data from Workshop 3.3 for the purposes of predicting whether traffic is "normal". Extend the workshop analysis in terms of "normal" vs "abnormal" traffic prediction, considering the following options: trying other clustering methods discussed in the block, other dimensionality reduction methods, and using cross validation to choose hyperparameters. Interpret your results.

Data Science Toolbox Portfolio Questions

04 Non-Parametrics and Missing Data

Daniel Lawson — University of Bristol

Block 4

Portfolio 04

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R04.1: How should we choose k in neighest neighbor methods? Consider the paper "OPTIMAL CHOICE OF k FOR k-NEAREST NEIGHBOR REGRESSION". What is the problem with using a fixed k for k-NN problems? Implement the method described in the paper and apply it to the example data from Lecture 4.1.

Question R04.2: Outlier removal using HDBSCAN was discussed in Lecture 4.2 and is recommended for the Machine Learning task of Topic modelling (which we interpret later in the course).

Use the example given in the documentation and implement a few choices for outlier removal. (It only requres `pip install bertopic` to run verbatim in colab). Present a figure summarising the impact of the outlier removal on the topic model, and discuss the results.

Question R04.3: Consider the Workshop 4.3 on Missing Data and in particular Section 5 on Evaluation, which asks how we know imputation worked, and specifically: *Q. How do we know that imputation has done a good job? Q. If the imputation went badly, how wrong could our estimates be?*

Continue the analysis and try to answer these questions, by performing an evaluation of the performance of the imputation procedure. To do this, you will need to leave further data out as cross valudation, or make adversarial imputation. Include your code and any non-essential figures as an appendix.

Data Science Toolbox Portfolio Questions

05 Supervised Learning and Ensembles

Daniel Lawson — University of Bristol

Block 5

Portfolio 05

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R05.1: The Kaggle competition MAP - Charting Student Math Misunderstandings is a classification problem based on free text information. Examine the code for a small variety of the top entries, and describe broadly how they have framed the problem. How does this practical classification differ from how it was presented in lectures, and how does it differ from the practical use case?

Question R05.2: Take a look at the Scikit-learn documentation on Histogram Boosting. Briefly summarise how it works, and "why it's faster". Giving a minimal code example as an appendix, compare its performance to "GradientBoostingClassifier" for a range of sample sizes. What is the trade-off between the two methods and what hyperparameter choices can impact it?

Question R05.3: Our Receiver-Operator Curve (ROC) is defined for binary classification problems. How can we extend it to multi-class classification problems? Using the package pROC in R or Multiclass Receiver-Operator curve in python, apply at least two multiclass scoring functions to the prediction of the `service` variable from which we derived the (binary) `http` variable for the data from Block 5 Workshop. Discuss how they differ both conceptually and in practice.

Data Science Toolbox Portfolio Questions

06 Decision Trees and Random Forests

Daniel Lawson — University of Bristol

Block 6

Portfolio 06

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R06.1: LightGBM Experiments show some impressive computational and accuracy results using "Best-first Decision Tree Learning", vs "Leaf-first" approaches. What empirical or theoretical evidence can you find (or create) to support or reject the claim that the critical difference is the "Best-first" approach?

Question R06.2: Decision trees align decision boundaries with **Features**. Either empirically or theoretically, discuss the use of using **PCA to construct features** for use in decision trees and random forests, boosted or otherwise. You can do this either by referencing the literature as a mini-review, or by extending the Block 06 workshop.

Question R06.3: Add LightGBM to the Block 06 workshop and compare its performance to Random Forests and xgBoost on at least one dataset from this course. With references to examples in the wild where Random Forests beat GBMs and vice versa, perform testing of selected hyperparameters to see if you can replicate these phenomena. From your examinations how confident are you of the superiority of one method over the other, and why?

Data Science Toolbox Portfolio Questions

07 Perceptrons and Neural Networks

Daniel Lawson — University of Bristol

Block 7

Portfolio 07

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R07.1: Investigate the Universal Approximation Theorem for MLPs, which says that an MLP with a single hidden layer of arbitrary width can approximate any continuous function.

Take `model_1`, the first MLP from lab code and modify its architecture. First train it with 1 hidden layer with 64 neurons, compare this with 2 hidden layers with 32 each etc. for a variety of different configuration of hidden layers (make sure you have activation functions between, you might want to use a loop inside your class to make this easier to define).

For a fixed number of epochs, record the training speed, model performance, minimum test loss etc. Produce plots (as an appendix) and explain your observations in the portfolio. You might want to watch this video for some inspiration.

Question R07.2: Improving CNN Performance. Augment the code for `model_2` (the CNN from the workshop) to improve performance by exploring hyper-parameters. Explore two of the following parameters (or others) to see what works best for performance. For each, write a short summary explaining what it does and document how it affects performance:

- Adjusting learning rate
- Use a different optimizer like Adam or AdamW
- Adjusting batch size
- Implement random transforms on your dataloaders (think about which transforms are appropriate for train/test loader)
- Dropout layers

Question R07.3: Read the paper "**Transformers need glasses!** Information over-squashing in language tasks" You could also listen to the podcast with the authors as well.

You're going to try and replicate some of the experiments and give an intuitive explanation of the results. To replicate the results, you should make a new google collab notebook and use some of the free transformer models. Within the collab notebook, remember to change runtime to GPU. You should use models from the transformers library for experiments (see example code at the bottom of the collab notebook). These are free but significantly worse performing than the models used in the paper so you should expect your results to reflect that!

Data Science Toolbox Portfolio Questions

08 Topic Model and LDA

Daniel Lawson — University of Bristol

Block 8

Portfolio 08

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R08.1: Consider the paper Exploring Topic Coherence over many models and many topics. Explain and motivate the models for coherence, and critique the conclusion that "LDA best learns descriptive topics while LSA is best at creating a compact semantic representation of documents and words in a corpus".

Question R08.2: In the workshop, LDAvis presented both "relevance" LDAvis: A method for visualizing and interpreting topics and "saliency" Termite: Visualization Techniques for Assessing Textual Topic Models. Examine the documentation for these terms and explain when one could be preferred over the other.

Question R08.3: From Eric Jang's "A Beginner's Guide to Variational Methods" or otherwise, explain what the KL Divergence between distributions is and how it relates to Variational Inference. What barriers are there to doing Variational inference in practice?

Data Science Toolbox Portfolio Questions

09 Algorithms for Data Science

Daniel Lawson — University of Bristol

Block 9

Portfolio 09

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R09.1: From Georgy Gimel'farb's Lecture or otherwise giving your sources, define, prove and explain with clear exposition and examples, the Algorithmic complexity of Quicksort in the average and worst case scenarios.

Question R09.2: For the Hash Table defined in Lecture 9.2, provide a clear explanation of both the average, amortized and worst case complexity for insertion. A formal proof is not required - focus instead on the clarity of exposition, for example, providing an appropriate figure.

Question R09.3: Extend the analysis of algorithmic complexity in Workshop 9.3 with other algorithms that we've discussed. Your code should be an appendix, and you should use your space for one figure, plus an explanation of what the algorithmic complexity is meant to be, and how it matches your experiments. It might be needed to extend the axes of the graph (and hence run longer) in order to see the patterns. Be careful with any parallelisation that might be silently being performed.

Data Science Toolbox Portfolio Questions

10 Parallel Algorithms

Daniel Lawson — University of Bristol

Block 10

Portfolio 10

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R10.1: By extending the benchmarking from Block 9 Workshop (09.3) to include parallel code as provided in Workshop 10.3, provide examples of *parallel speedup* in which a) the *efficiency* is 1, and b) the efficiency is lower than 1 but still of value (i.e. the parallel algorithm does more overall compute than the sequential but is quicker). These should be algorithms for which the *computational efficiency* exhibits these features - they may have constant terms that make practice harder. Focus your writeup on the choice and scaling of the algorithms.

Question R10.2: Investigate Spark (e.g. using pyspark or sparkR) and implement a simple mapping-and-reducing problem, providing the code as an appendix and writing up in the format of a tutorial.

Question R10.3: Explain the difference between Matrix Multiplication as implemented on a CPU vs a massively parallel GPU, from the paper Understanding the Efficiency of GPU Algorithms for Matrix-Matrix Multiplication. In terms of concepts we've covered in DST, what is the take-home message?

Data Science Toolbox Portfolio Questions

11 Ethics and Privacy

Daniel Lawson — University of Bristol

Block 11

Portfolio 11

Choose **one question** and write up to **one page** about it. You are free to conduct further experiments to add weight to your results, and any additional material you generate can be submitted as an appendix. See The Assessment Page for advice.

These questions may make reference to the content from the current block.

Question R11.1: For a dataset that you have explored elsewhere, or otherwise, investigate whether you are able to re-identify the data without labels. The portfolio should be about the general ideas for how anonymisation and re-identification work, which are illustrated through case studies (i.e. discussion of specific examples). You are not expected to undertake a practical analysis but any coding you do to support your claims belong in an appendix.

Question R11.2: Briefly explain the goal of differential privacy. Compare and contrast Stochastic Gradient Descent and its differentially private counterpart, Algorithm 1 from Deep Learning with Differential Privacy, relating it to one aspect of the course (spanning lecture notes, coursework you've done, or the portfolio).

Question R11.3: Consider Section 3.2 Measures of Algorithmic Bias of Algorithmic Fairness. With an example (from your own work or from the paper) contrast **Equal opportunity** and one other measure, explaining how they could be used in practice.

Marking Criteria

The mark ranges and descriptions in normal type below are the University of Bristol Generic Marking criteria that apply to any assessment at the University - these can be found at https://www.bristol.ac.uk/academic-quality. The descriptions in bold type are additional maths-specific criteria introduced primarily to clarify the descriptors in the case of marking maths examinations.

0-100 scale	**Criteria to be satisfied** University generic marking criteria in normal type, **Maths-specific marking criteria in bold**
100 94 89	• Work would be worthy of dissemination under appropriate conditions • Mastery of advanced methods and techniques at a level beyond that explicitly taught • Ability to synthesise and employ in an original way ideas from across the subject • In group work, there is evidence of an outstanding individual contribution • Excellent presentation • Outstanding command of critical analysis and judgement **and** • **Work develops concepts not directly presented in course material or uses known concepts to answer hard, unfamiliar questions that require calculations/methods not similar to any course material** • **An elegance of mathematical work beyond that expected for the level of the course** • **Of a quality that could be distributed to fellow students as an example of exceptional work**
83 78 72	• Excellent range and depth of attainment of intended learning outcomes • Mastery of a wide range of methods and techniques • Evidence of study and originality clearly beyond the bounds of what has been taught • In group work, there is evidence of an excellent individual contribution • Excellent presentation **and** • **On standard but unfamiliar problems, carrying out calculations with no errors of understanding** • **Demonstrates a high level of technical competence with very few mistakes of any kind** • **Great clarity in mathematical arguments**
68 65 62	• Attained all the intended learning outcomes • Able to use well a range of methods and techniques to come to conclusions • Evidence of study, comprehension and synthesis beyond the bounds of what has been explicitly taught • Very good presentation of material • Able to employ critical analysis and judgement • Where group work is involved there is evidence of a productive individual contribution **and** • **Able to make a good attempt at standard but unfamiliar problems, with some minor errors** • **Demonstrates technical competence, perhaps with some shortcomings** • **Clear mathematical arguments**

0-100 scale	Criteria to be satisfied University generic marking criteria in normal type, **Maths-specific marking criteria in bold**
58 55 52	• Some limitations in attainment of learning objectives, but has managed to grasp most of them • Able to use most of the methods and techniques taught • Evidence of study and comprehension of what has been taught • Adequate presentation of material • Some grasp of issues and concepts underlying the techniques and material taught • Where group work is involved there is evidence of a positive individual contribution **and** • **Able to start standard but unfamiliar problems but with significant errors** • **Able to complete competently "bookwork" questions that have been seen in the course material**
48 45 42	• Limited attainment of intended learning outcomes • Able to use a proportion of the basic methods and techniques taught • Evidence of study and comprehension of what has been taught, but grasp insecure • Poorly presented • Some grasp of the issues and concepts underlying the techniques and material taught, but weak and incomplete **and** • **Able to complete "bookwork" questions that have been seen in course material with few errors** • **Gaps or inconsistencies in the mathematical argument**
35	• Attainment of only a minority of the learning outcomes • Able to demonstrate a clear but limited use of some of the basic methods and techniques taught • Weak and incomplete grasp of what has been taught • Deficient understanding of the issues and concepts underlying the techniques and material taught **and** • **Able to reproduce work seen in course material, but with some errors**
7-29	• Attainment of nearly all the intended learning outcomes deficient • Lack of ability to use at all or the right methods and techniques taught • Inadequately and incoherently presented • Wholly deficient grasp of what has been taught • Lack of understanding of the issues and concepts underlying the techniques and material taught **and** • **Unable to reproduce satisfactorily even "bookwork" questions that have been seen in course material**
0	• No significant assessable material, absent or assessment missing a "must pass" component