Data Science Toolbox Portfolio Questions

Long form questions

Daniel Lawson — University of Bristol

due Week 13

Part of Data Science Toolbox.

Deadline: Wednesday Noon, Week 13 (January)

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.
  - Reflection: Deeper descriptions of material you have examined during the Worksheets and Workshop.
- The Reflection content should be no more than 1 page per block.

Portfolio Content

Reflection 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 on each block from 2-11. Block 1 is marked similarly but is formative, i.e. does not contribute to your mark. The deadline is the start of TB2. In each block you will do two activities:

1. Multiple choice questions submitted via Noteable (log in via Blackboard). These should be straightforward, either direct from your notes or with very 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 each Block. Therefore the whole Portfolio is only 10 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 15 pages** for the portfolio content, with answers that are clearly too long being be 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:

Follow. 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: Imagine that you are tasked with making a temperature prediction for 2040 based on the Temperature Data used in Workshop 2.3. Design 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.

Question R02.2: It was claimed without proof 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 R02.3: Consider the final non-linear stepwise model that was obtained for the diamond data (the object called `modelcvintstep` and named `intstep`). It has the highest R^2 with the test data, and highest AIC of all models considered. Investigate and discuss the ways that this model may be considered *best* and how it may yet be bettered by other models considering the same model space and data (i.e. all pairwise quantitative features plus the ordinal factors). Discuss what interpretation we can make on the linear and non-linear effects of the parameters.

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: Imagine that you are trying to understand the Cyber Security data from Workshop 3.3 for the purposes of predicting whether traffic is "normal". Consider the advantages and disadvantages of enriching the feature set via a) dimensionality reduction, and b) clustering, for the purpose of passing to a classifier. You may wish to perform experiments (and cite results placed in your appendix) for this task.

Question R03.2: Describe the vanilla UPGMA (Average Linkage Clustering) algorithm and compare it to an efficient and more scalable approach, for example Sparse UPGMA, paying specific attention to how it can be made more efficient than $O(N^3)$.

Question R03.3: 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?".

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: Conduct a Nearest Neighbour imputation of the dataset presented in Workshop 4.3, plus any other imputation that you wish to try, and document it as part of your appendix. Then consider that you are tasked with "tidying up" this dataset for onwards analysis as part of an anonymised health data competition. What do you choose to present and why?

Question R04.2: Consider the paper "Kernel Methods in Machine Learning". Write a simple explanation suitable for Masters' level class in Data Science describing how a **polynomial kernel** would be used in this context, and explain its use case.

Question R04.3: Run the analysis of missing data described in the finalfit vignette from which our `colon` dataset came. Contrast their findings to those in the workshop. Make sure to document any work you do beyond the verbatim code, e.g. if you run their analysis with more variables or ours with fewer.

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: Imagine that you are a data-science consultant who has provided the analysis in Block 5 Workshop to a company. Give a 1-page "executive summary" of your conclusions, being sure to address: a) predictive performance, b) computational concerns, and c) interpretation. You are welcome to include additional analyses in the appendix, and reference figures, as appropriate.

Question R05.2: Your boss tells you to perform Bagging as it will improve performance. Explain what bagging can do for you, as well as identifying the limitations and costs that it has.

Question R05.3: Read a paper on the Quadratic Programming problem for SVM's such as Multiplicative Updates for Nonnegative Quadratic Programming in Support Vector Machines to understand and summarise quadratic programming, and explain one method for finding a solution to it. What problems can Quadratic Programming solve in machine learning generally, including but not limited to SVMs? What can it not handle? Emphasise the challenges that the quadratic component creates over simpler problems.

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: Read Section 2 of XGBoost: A Scalable Tree Boosting System. Explain what the second-order gradient statistics for each leaf are, and how and why they are used in boosting.

Question R06.2: Extend the workshop to make a specific comparison between the choice of splitting measure - i.e. compare the **Gini index** to **ID3** and potentially other measures, both for **Decision Trees** and inside a tree ensemble, either bagged decision trees or Random Forest. You may find Section 2.1 of Decision trees: a recent overview helpful. Your code and figures would be an appendix, you should focus on any conceptual issues and conclusions in your portfolio.

Question R06.3: 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. You can do this either by referencing the literature as a mini-review, or by extending the workshop as in Q6.2.

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: Read Section 5 on Layer-wise relevance propagation (LRP) from Methods for interpreting and understanding deep neural networks. Explain what LRP is, and how it extracts interpretable features. Potentially referencing the keras-explain manual, discuss any practical constraints.

Question R07.2: Consider the paper On Calibration of Modern Neural Networks, paying particular attention to "Temperature scaling". Explain this, by a) explaining what the calibration problem of neural networks is, b) explaining how we know a model is calibrated, and c) how temperature scaling addresses this problem.

Question R07.3: In The Tradeoffs of Large Scale Learning it is shown that if we take into account approximation error, estimation error and optimization error, then Stochastic Gradient Descent can be seen to converge faster in terms of compute cost than regular Gradient Descent (Table 2). Describe just what is needed to interpret the key results for GD and SGD, for "small scale" and "large scale learning" and briefly interpret in terms of learning neural networks.

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 www.bristol.ac.uk/esu/assessment/codeonline.html. 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