Portfolio B: Data Science and Engineering

Part of Data Science Toolbox. Due Week 24

Data Science Portfolios Overview

- Each is worth 20% of the course
- Submit two portfolios, one for each term
- Individually assessed, follow UoB coursework assessment guidelines

What is a portfolio?

- A portfolio in Data Science Toolbox is an opportunity to delve deeper into some data science subjects of your choice.
- It consists of:
  1. **Context** linking to your course notes and other work,
  2. A **Topic Note**,
  3. **Reflection**.
- You will write a **Topic Note** about one **Topic**. These are short and approachable **technical notes** that explain and reflect on this topic at a level suitable for your peers.
  - Make it useful for you to return back to in 5 years time. It therefore needs to be well linked to your course material, well referenced, and easily readable.
  - You will base your work on a paper, book chapter or high-quality technical note that extends content in our lecture notes.
  - You do not need to work with the whole paper, only the parts that pertain to the block.
  - The level of detail and content expected is appropriate for a high-quality technical blog post. You are welcome to make them into actual blog posts to make a public portfolio.
- The topics should be taken from the content blocks:
  - No more than one topic per block.
  - Topics must come from the same semester (Blocks 1-6 for Portfolio A or 7-12 for Portfolio B).
- You can choose the degree to which you adopt a mathematical vs an experimental code-based data-science perspective. The level of mathematical detail should not be significantly less than our lecture notes, unless the paper is very conceptual or applied.
- The topic should extend the course notes by a small amount.
- We provide example suitable papers but you are free to pursue others:
  - If you choose a paper off-list, confirm its appropriateness before doing too much work. You can always use it as a reference if it is deemed unsuitable.
  - Papers may be inappropriate if they are too easy, too hard, or too far from the topic.
  - Many technical details, especially for machine learning or older sub-

jects, are given in non-academic places. Discussing these is fine if they contain enough technical depth. You would still need to appropriately reference the primary literature.

- Whilst it will be necessary to read around the subject, you are only expected to read only the target paper in depth.
  - Typically you would be expected to find (and have skimmed and read the abstract of) a few important papers in the area.
- You are advised to use the headings below.

Content for Context and Reflection

Your introduction and discussion are where you link the Topic Notes to your working. Suitable content includes:

- **Context**: (4 pages) Link the papers directly into the course content, preferably where possible with mathematical reasoning. Explain the relevant course content in your own words. What is the problem that your topic is going to go on to address, that the course content is not adequate for?
- **Reflection**: (2 pages) Why did you choose these topics? How useful is this likely to be in practice for data science problems? How would you use (or could you use) them in a practical?

If your topics do not feel connected, then you may split the context and reflection for each topic, rather than combining them.

Suggested structure of a Topic Note

Total: approx 7 pages (longer if you use larger or many figures.)

1. **Background**: (approx 2 pages) Explain what the method is trying to solve, give an overview of the method, and briefly discuss its origins, history and purpose in broad terms It is likely to be appropriate to discuss alternative approaches to the same problem, and more modern approaches to the same issue.
2. **Methods**: (approx 2 pages) Explain the method in some mathematical detail.
3. **Usage**: (approx 2 pages) How does the method get applied in practice? What issues arise? Discuss this either with the aid of a real-data or conceptual example. It would be common to re-use an example from the paper.
4. **Summary**: (approx 1 page) Wrap everything up neatly and point to any further reading. Any criticisms of the paper or approach?

Assessment Details:

The assessment is 60% Topic 1 and 40% Context and Reflection.

- **Topics** and **Context and Reflection** are assessed equally between:

  - Breadth and Clarity.
    * *Your description should be accessible and do a good job of placing your topic in context. It should demonstrate breadth of understanding beyond the narrow details of the topic. It should build upon appropriate resources.*
  - Depth of Insight.
    * *You should try to bring additional insight over the content, by appropriate examination of mathematical or programming concerns. This is* **not** *undertaking additional mathematics or data analysis, but explaining and expanding what is there.*
  - Literature and Understanding.
    * *Cite your sources, build up a repertoire of useful content. Link to analogous problems and resources. Show that you understand both them and the underlying lecture note content.*
  - Structure and Description.
    * *Your content should be well introduced, and easy to read and understand. Make good figures and explain them. Structure your project well, stick to the point.*

Portfolio A topic suggestions

Block 7: Topic Models and Bayes

- Bag of Words:
  - Python Bag of Words: p259 Python Machine Learning (Raschka & Mirjalili, 2nd ed 2017).
  - Topic Modeling and Latent Dirichlet Allocation: An Overview (Weifeng Li, Sagar Samtani and Hsinchun Chen)
  - Stephen Robinson, Microsoft Research Understanding Inverse Document Frequency: On theoretical arguments for IDF
- LDA
  - B. Barde and A. Bainwad. "An overview of topic modeling methods and tools" (2017) ICICCS 745-750.
  - Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation", Journal of machine Learning research 3.Jan (2003): 993-1022.
- Variational Inference: * Blei, Kucukelbir and McAuliffe. "Variational Inference: A Review for Statisticians", JASA (2017): 859-877. * Blei and Jordan. "Variational inference for Dirichlet process mixtures", Bayesian analysis 1.1 (2006): 121-143. * A Beginner's Guide to Variational Methods, by Eric Jang.
- Cutting edge topics, for example:
  - How do text prediction work?
  - How does ChatGPT work? (link to Neural Network content, may cover both topics. . .)

Block 8: Algorithms for Data Science

- Algorithms for Big Data:
  - Many topics from Open DSA - Data Structures and Algorithms would be appropriate.
  - Broder & Mitzenmacher "Network Applications of Bloom Filters: A Survey" (2003) Internet Mathematics 1:485-509
  - Geravand & Ahmadi "Bloom filter applications in network security: A state-of-the-art survey" (2013) Computer Networks 57:4047-4064
  - Goyal, Daume & Cormode "Sketch Algorithms for Estimating Point Queries in NLP" (2012) Proc. EMNLP.

Block 9: Perceptions and Neural Networks

- Book chapter content: Explore:
  - Chapter 11 of The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Friedman, Hastie and Tibshirani).
  - Russell and Norvig Artificial Intelligence: A Modern Approach
  - Chapter 20 Section 5: Neural Networks
- Theoretical practicalities:
  - Bengio 2012 Practical Recommendations for Gradient-Based Training of Deep Architectures (in the book "Neural Networks: Tricks of the Trade")
  - Kull et al 2019 NeurIPS Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with Dirichlet calibration
  - Swish: Ramachandran, Zoph and Le Searching for Activation Functions
- Cutting edge topics, for example:
  - How does AlphaZero work?
  - What is an adversarial Neural Network?

Block 10: Parallel Algorithms

- Chapter 27 of Cormen et al 2010 Introduction to Algorithms
- Real-world questions:
  - How is multi-threading handled in routine tasks? How does it vary between the operating system, web browser, games, consoles, mobiles? Could we do scientific computing on them?
  - What are the scientific computing options for vectorisation? What can be automated? e.g. Numpy's multi_dot
  - MapReduce algorithm for matrix multiplication
- A Brief Overview of Parallel Algorithms

Block 11: Parallel Infrastructure and Spark

- We don't go into lots of detail on this, so even something simple is an extension of our content.

- General parallel algorithms:
  - Streaming and Sketching
  - Parallel algorithms for dense matrix multiplication
- Write yourself a Python-to-Spark comparison, following e.g.:
  - Pandas vs Pyspark
  - Spark RDD guide
  - pyspark
  - Spark RDD Transformations

Block 12: Ethics and Privacy

- Laws governing data science:
  - Human Rights Act (HRA 1998)
  - EU General Data Protection Regulation (GDPR 2018)
  - Data Protection Act 2018 (DPA 2018)
- Privacy:
  - The Algorithmic Foundations of Differential Privacy by Dwork and Roth (2014).
  - ONS policy on disclosure control.
  - Sweeney 1997. Weaving technology and policy together to maintain confi-dentiality. Journal of Law, Medicines Ethics, 25:98–110.
  - Narayanan and Shmatikov 2008. Robust de-anonymization of largesparse datasets (how to break anonymity of the netflix prize dataset). IEEE Sec. and Priv.
  - Statistical Disclosure Attacks by George Danezis.
- Interpretability and fairness:
  - Book: "Interpretable Machine Learning: A Guide for Making Black Box Models Explainable." Christoph Molnar 2019.
  - Algorithmic Bias Tutorial by Francesco Bonchi with Slides from KDD 2016
  - Hardt, Price and Srebo Equality of Opportunity in Supervised Learning 2016 explored in https://blog.acolyer.org/2018/05/07/equality-of-opportunity-in-supervised-learning/.