Portfolio A: Statistical Machine Learning

Part of Data Science Toolbox. Due Week 13

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. Two **Topic Notes**,
  3. **Reflection**.
- You will write a **Topic Note** about two **Topics**. These are short and approachable **technical notes** that explain and reflect on this topic at a level suitable for your peers.
  - 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.
  - All 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.
- There does not need to be an explicit cyber-security perspective in the paper.
- 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 subjects, 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 would not normally be expected to cite the Data Science Toolbox course content, unless it was the most appropriate source for explaining some content.

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. How are the two Topics related? How do they each fit in with and build upon the course notes? How have you used (or could you use) them in a practical? * **Reflection**: (2 pages) Why did you choose these topics? How useful is this likely to be in practice for data science problems?

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 may also 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?

Assessment Details:

The assessment is 35% Topic 1, 35% Topic 2, and 30% 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 1: Exploratory Data Analysis

- This block is hard to find high-quality resources on. I would instead recommend choosing a topic from below, and undertaking an EDA activity to support that understanding. You are welcome to find suggestions however. Following the data trail for an excellent vizualisation could also count, e.g. https://visme.co/blog/best-data-visualizations/, as could visualising MITRE ATT&CK as done e.g. on Andy Applebaum's Medium blog. However you have to find and provide technical content.

Block 2: Regression and Statistical Testing

- Cross Validation:
  - Prediction error estimation: a comparison of resampling methods by A. Molinary, R. Simon and R. Pfeiffer (Bioinformatics, 2005). *You could update the review with modern methods, or look in detail at the mathematics of one of them.*
- Regression:
  - Robust Regression by K. Lawrence and J. Arthur. *This book contains a number of contributed chapters, most of which would be suitable.*

Block 3: Latent Structures, PCA, and Clustering

- PCA:
  - Principal component analysis by H. Abdi and L. Williams. *It would be natural to consider the difference between PCA, Correspondance analysis and factor analysis.*
  - Robust PCA and classification in biosciences by M. Hubert and S. Engelen. *You could explain what Robust PCA is, why we want it, and other methods that attempt to do the same thing.*
- Clustering
  - Theoretical guarantees for EM under misspecified Gaussian mixture models by R. Dwivedi, K. Khamaru, M. Wainwritght, and M. Jordan,

It would be natural to focus on understanding the EM-algorithm and how mis-specification is handled.
  - GMCM: Unsupervised Clustering and Meta-Analysis Using Gaussian Mixture Copula Models by A. Bilgrau et al. *You could explain what a copula is and why we want one. Discuss how a GMM is learned.*

Block 4: Non-parametrics and Missing Data

- Non-Parametric statistics
  - Outlier Detection: Methods, Models, and Classification by A. Boukerche, L. Zheng and O. Alfandi. *As it is an already accessible review, you could drill into the examples with a tutorial-style code, or explain one method with some mathematical detail, to add insight.*
  - Kernel Methods in Machine Learning by Hofmann, Schoelkopf, & Smola (2008). *A relatively dense paper, so explaining with examples would add insight.*
- Missing data:
  - Missing Data, Imputation, and the Bootstrap B. Efron. *Explore the bootstrap and missing-data through an example.*

Block 5: Supervised Learning and Ensembles

- The Relationship Between Precision-Recall and ROC Curves by J. Davis and M. Goadrick. *How should we use ROC and PR curves? What theory can be said about them?*
- Tree Boosting With XGBoost: Why Does XGBoost Win "Every" Machine Learning Competition? by D. Nilsen. *This is a thesis, you can try to summarise it down to the core components and focus on the interesting conclusion.*
- Bagging, boosting and stacking in machine learning which is a Stack Exchange question. *You could find primary literature to answer this question in an academically structured way.*

06 Decision Trees and Random Forests

- A working guide to boosted regression trees by J. Elith, J. Leathwick, and T. Hastie. *You could replicate with a different example, bring the guide up-to-date, or find the mathematical explanation of boosted regression trees.*
- Understanding variable importances in forests of randomized treesby G. Louppe, L. Wehenkel, A. Sutera, P. Geurts. *You could explain this in plain language with an example*