TÉCNICO LISBOA

Mathematics for Machine Learning

2021-2022

Project 3 - A deeper understanding

In this course you learned something (we hope) about several mathematical frameworks with fundamental application to machine learning. The subject matter we covered is never-the-less just a slice of the most fundamental ideias. Due to time constraints, there are many beautiful and useful mathematical theories that we did not cover, and there are also deeper results or applications in the subject matter we did cover.

In this project you have the opportunity to choose a theme you are interested in for a deeper look and understanding. Below are several suggestions, but you might also suggest a theme of your own. For this project you can use Python or R, as suits the theme better.

You have to deliver by the deadline of **June 20th** an article-type 8 page report including explanation of what you did, graphics, analysis, and discussion of your results, together with the commented code of your project. You must also make a 10 minute presentation to the Class (this will happen on the lectures of June 20th and 21st). This assignment is **by Groups of 2**.

List of themes

1- **Interval Principal Component Analysis on Process Monitorin**g:

Tarek Ait-Izem, M-Faouzi Harkat, Messaoud Djeghaba, Frédéric Kratz, On the application of interval PCA to process monitoring: A robust strategy for sensor FDI with new efficient control statistics, *Journal of Process Control*, **63**, 29-46, 2018. https://doi.org/10.1016/j.jprocont.2018.01.006

Tasks:
- Read the paper and learn about Interval PCA.
- Replicate the simulation study presented in the paper.
- Find a dataset or (simulate one based on the real example discussed in the paper) and analyze it. Compare your results with a non-symbolic approach.
- Discuss and do a critical analysis of your findings.

Further reading:
- Margarida Azeitona, Classical and Robust Symbolic Principal Component Analysis for Interval Data, Master Thesis, 2015. https://fenix.tecnico.ulisboa.pt/cursos/mma/dissertacao/1972678479052893
- Rodrigo Serrão, Symbolic Formulation for Principal Component Analysis of Interval Valued Data, Master Thesis, 2021. https://fenix.tecnico.ulisboa.pt/cursos/mma/dissertacao/1691203502344165

2- **Optimized Dimensionality Reduction Methods for Interval-Valued Variables and Their Application to Facial Recognition**

J. Arce Garro and O. Rodríguez Rojas. Optimized Dimensionality Reduction Methods for Interval-Valued Variables and Their Application to Facial Recognition. *Entropy*, ***21***(10), 1016, **2019.**
https://doi.org/10.3390/e21101016

Tasks:
- Read the paper and learn about Interval PCA.
- Design a simulation study to compare the performance of the proposed method with other already presented in the literature.
- Replicate the analysis of the example explored in the paper. Compare your results with a non-symbolic approach and other interval PCA methods.
- Discuss and do a critical analysis of your findings.

Further reading:
- Margarida Azeitona, Classical and Robust Symbolic Principal Component Analysis for Interval Data, Master Thesis, 2015. https://fenix.tecnico.ulisboa.pt/cursos/mma/dissertacao/1972678479052893
- Rodrigo Serrão, Symbolic Formulation for Principal Component Analysis of Interval Valued Data, Master Thesis, 2021. https://fenix.tecnico.ulisboa.pt/cursos/mma/dissertacao/1691203502344165

3- ROBPCA: A New Approach to Robust Principal Component Analysis

Mia Hubert and Peter J. Rousseeuw. ROBPCA: A New Approach to Robust Principal Component Analysis. *Technometrics, 47(1), 2005, 64-79.*
https://www.jstor.org/stable/25470935

Tasks:
- Read the paper and learn about robust PCA and outlier detection.
- Replicate one scenario of the simulation study.

- Find a dataset and analçyse it using what you have learned. Compare your results with non-robust approaches.
- Discuss and do a critical analysis of your findings.

Further reading:
A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining and Knowledge Discovery. 8: null. doi: 10.1002/widm.1280.
https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1280

4- Robust statistics for outlier detection

Peter J. Rousseeuw and Mia Hubert Robust statistics for outlier detection. *WIREs Data Mining Knowl Discov*, 8:e1236. doi: 10.1002/widm.1236, 2018.
https://wis.kuleuven.be/stat/robust/papers/publications-2018/rousseeuwhubert-anomalydetection-wires-dmkd-2018.pdf

Tasks:
- Read the paper and learn about robust PCA and robust Mahalanobis distance as outlier detection methods.
- Design a small simulation study to compare classical and robust methods.
- Find a dataset and analyse it using what you have learned. Compare your results with non-robust approaches.
- Discuss and do a critical analysis of your findings.

Further reading:
A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining and Knowledge Discovery. 8: null. doi: 10.1002/widm.1280.
https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1280

5- Robust statistics for outlier detection

Chen, X., Zhang, B., Wang, T. *et al.* Robust principal component analysis for accurate outlier sample detection in RNA-Seq data. *BMC Bioinformatics* 21, 269 (2020).
https://doi.org/10.1186/s12859-020-03608-0

Tasks:
- Read the paper and learn about robust PCA as outlier detection methods.
- Design a small simulation study to compare classical and robust methods.

- Find a dataset and analyse it using what you have learned. Compare your results with non-robust approaches.
- Discuss and do a critical analysis of your findings.

Further reading:
A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining and Knowledge Discovery. 8: null. doi: 10.1002/widm.1280.
https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1280

6- Local projections for high-dimensional outlier detection

T. Ortner, P. Filzmoser, M. Rohm, S. Brodinova, and C. Breiteneder. Local projections for high-dimensional outlier detection. Metron, 79:189-206, 2021. doi.org/10.1007/s40300-020-00183-5.
https://link.springer.com/article/10.1007/s40300-020-00183-5

Tasks:
- Read the paper and learn about robust outlier detection methods.
- Replicate a part of the paper's simulation study to compare classical and robust methods.
- Find a dataset and analyse it using what you have learned. Compare your results with non-robust approaches.
- Discuss and do a critical analysis of your findings.

Further reading:
A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining and Knowledge Discovery. 8: null. doi: 10.1002/widm.1280.
https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1280

7- Outlier detection in interval data
A.P. Duarte Silva, P. Filzmoser, and P. Brito. Outlier detection in interval data. Advances in Data Analysis and Classification, 12(3), 785-822, 2018.
https://link.springer.com/article/10.1007/s11634-017-0305-y

Tasks:
- Read the paper and learn about robust outlier detection methods.
- Replicate a part of the paper's simulation study to compare classical and robust methods.
- Find a dataset and analyse it using what you have learned. Compare your results with non-robust approaches.

- Discuss and do a critical analysis of your findings.

Further reading:
A. Zimek and P. Filzmoser. There and back again: Outlier detection between statistical reasoning and data mining algorithms. WIREs Data Mining and Knowledge Discovery. 8: null. doi: 10.1002/widm.1280.
https://wires.onlinelibrary.wiley.com/doi/epdf/10.1002/widm.1280

8- Word Embeddings and NLP
Kevin Lund and Curt Burgess (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instrumentation, and Computers, 28:203–208.
https://link.springer.com/content/pdf/10.3758/BF03204766.pdf
Mikolov, T., Chen, K., Corrado, G.S., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*.
https://arxiv.org/abs/1301.3781
Jeffrey Pennington, Richard Socher, and Christopher Manning (2014) GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543
https://nlp.stanford.edu/pubs/glove.pdf

Tasks:
- Read the papers and learn about the idea and the various techniques for creating word embeddings.
- Implementation at least two of the described algorithms.
- Find a (small) text dataset (if possible non-english, but in your native language). Process the embeddings, and visualize the semantic space.
- Your report should do a critical comparison of the different embedding techniques and methods, illustrated with your example.

9- Support Vector Machines and Kernel Methods
Bishop (2006). Pattern Recognition and Machine Learning, Chapters 6 and 7.
https://www.microsoft.com/en-us/research/publication/pattern-recognition-machine-learning/
Shalev-Shwartz, Ben-David (2014). Understanding Machine Learning, Chapters 15 and 16
http://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning

Tasks:
- Read the book chapters and learn about Kernel Methods and SVMs.
- Find a classification dataset, implement and use the algorithms to analyze and classify the data.

- Your report should present the theory and describe its application to the analyzed dataset.

10- Approximating functions with Deep vs Shalow Neural Networks

Telgarsky, M.. (2016). Benefits of depth in neural networks. 29th Annual Conference on Learning Theory, in Proceedings of Machine Learning Research 49:1517-1539
https://proceedings.mlr.press/v49/telgarsky16.html
D. Rolnick, M.Tegmark (2018), The power of deeper networks for expressing natural functions, International Conference on Learning Representations (ICLR) 2018
https://openreview.net/pdf?id=SyProzZAW

Tasks:

- Read the papers.
- Define a set of functions with different characteristics, and approximate them with different NN architectures (everything should be programmed by you, only the basic python packages (E.g. numpy) should be used).
- Your report should resume the theory and critically analyze the experiments.

Further reading:

Roberts, Hanin (2022).The Principles of Deep Learning Theory, Cambridge University Press.
https://arxiv.org/abs/2106.10165