UNIVERSITY OF EASTERN FINLAND

Please carefully read and follow the general instructions regarding coding assignments. Failing to meet the requirements might lead to penalties. https://elearn.uef.fi/mod/page/view.php?id=248672

If you suspect that something is wrong with some task instructions, please contact the lecturer.

If you face persistent issues while working on a task, do ask for help, e.g. during a course meeting or by contacting the lecturer via email.

Choose one of the research papers listed below, implement and test the main proposed algorithm.

The report should start with an introduction of the algorithm. Consider that your audience consists of other students from the course, who know about the related concepts studied in the course but are not familiar with that particular algorithm.

Paper 1. Datar, M., Gionis, A., Indyk, P., and Motwani, R. (2002). *Maintaining stream statistics over sliding windows.* SIAM journal on computing, 31(6), pp. 1794–1813.

Complementary material:

- Section 4.6 of the textbook Leskovec, J., Rajaraman, A., and Ullman, J. D. (2014). *Mining of massive datasets.* Cambridge university press. http://www.mmds.org/#ver21

Generate synthetic data, i.e. artificially generated binary sequences, to test your implementation.

Paper 2. Domingos, P., and Hulten, G. (2000). *Mining high-speed data streams.* In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 71–80).

Complementary material:

- Hulten, G., Spencer, L., and Domingos, P. (2001). *Mining time-changing data streams.* In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 97–106.
- Bifet, A., and Kirkby, R. (2009). *Data stream mining a practical approach.* https://www.cs.waikato.ac.nz/~abifet/MOA/StreamMining.pdf

Use datasets from the *UCI Machine Learning repository* (https://archive.ics.uci.edu/, e.g. `Spambase`, `Post Operative Patient`, `Adult` datasets), to test your implementation.

Paper 3. Liu, F. T., Ting, K. M., and Zhou, Z. H. (2008). *Isolation forest.* In Proceedings of the eighth IEEE International Conference on Data Mining, pp. 413–422.

Complementary material:

- Liu, F. T., Ting, K. M., and Zhou, Z. H. (2012). *Isolation-based anomaly detection.* ACM Transactions on Knowledge Discovery from Data (TKDD), 6(1), 3.

Use datasets from the *UCI Machine Learning repository* (https://archive.ics.uci.edu/, e.g. `Arrhythmia`, `Thyroid Disease`, `Shuttle` datasets), to test your implementation.

Paper 4. Keogh, E., Lin, J., Lee, S. H., and Van Herle, H. (2007). *Finding the most unusual time series subsequence: algorithms and applications.* Knowledge and Information Systems, 11(1), pp. 1–27.

Complementary material:

- Keogh, E., Lin, J., and Fu, A. (2004). *HOT SAX: Finding the most unusual time series subsequence: Algorithms and applications.* In Proceedings of the fifth IEEE International Conference on Data Mining, pp. 440–449.
- Patel, P., Keogh, E., Lin, J., and Lonardi, S. (2002). *Mining motifs in massive time series databases.* In Proceedings of the second IEEE International Conference on Data Mining, pp. 370–377.

Use time-series from the paper's webpage (http://www.cs.ucr.edu/~eamonn/discords/) to test your implementation.