The Journal of Open Source Software

PyNomaly: Anomaly detection using Local Outlier Probabilities (LoOP).

Valentino Constantinou[1]

1 NASA Jet Propulsion Laboratory

DOI: 10.21105/joss.00730

Software

- Review
- Repository
- Archive

Submitted: 08 May 2018
Published: 15 May 2018

Summary

`PyNomaly` is a Python 3 implementation of LoOP (Local Outlier Probabilities) (Kriegel et al. 2009). LoOP is a local density based outlier detection method by Kriegel, Kröger, Schubert, and Zimek which provides outlier scores in the range of [0,1] that are directly interpretable as the probability of a sample being an outlier. `PyNomaly` also implements a modified approach to LoOP introduced by Hamlet et. al., which may be used for applications involving streaming data or where rapid calculations may be necessary.

The outlier score of each sample is called the Local Outlier Probability. It measures the local deviation of density of a given sample with respect to its neighbors as Local Outlier Factor (LOF) (Breunig et al. 2000), but provides normalized outlier scores in the range [0,1]. These outlier scores are directly interpretable as a probability of an object being an outlier. Since Local Outlier Probabilities provides scores in the range [0,1], practitioners are free to interpret the results according to the application.

Like LOF, it is local in that the anomaly score depends on how isolated the sample is with respect to the surrounding neighborhood. Locality is given by k-nearest neighbors, whose distance is used to estimate the local density. By comparing the local density of a sample to the local densities of its neighbors, one can identify samples that lie in regions of lower density compared to their neighbors and thus identify samples that may be outliers according to their Local Outlier Probability.

`PyNomaly` includes an optional *cluster_labels* parameter. This is useful in cases where regions of varying density occur within the same set of data. When using *cluster_labels*, the Local Outlier Probability of a sample is calculated with respect to its cluster assignment. In addition, PyNomaly also includes a modification of LoOP introduced by Hamlet et. al. (Hamlet et al. 2017) that could be used in applications of streaming data.

Research

PyNomaly is currently being used in the following research:

- Y. Zhao and M.K. Hryniewicki, "XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning," International Joint Conference on Neural Networks (IJCNN), IEEE, 2018.

Acknowledgements

The authors recognize the support of Kyle Hundman and Ian Colwell.

References

Breunig, Markus M., Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. "LOF: Identifying Density-Based Local Outliers." In *Proceedings of the 2000 Acm Sigmod International Conference on Management of Data*, 93–104. SIGMOD '00. New York, NY, USA: ACM. https://doi.org/10.1145/342009.335388.

Hamlet, Connor, Jeremy Straub, Matthew Russell, and Scott Kerlin. 2017. "An Incremental and Approximate Local Outlier Probability Algorithm for Intrusion Detection and Its Evaluation." *Journal of Cyber Security Technology* 1 (2):75–87. https://doi.org/10.1080/23742917.2016.1226651.

Kriegel, Hans-Peter, Peer Kröger, Erich Schubert, and Arthur Zimek. 2009. "LoOP: Local Outlier Probabilities." In *Proceedings of the 18th Acm Conference on Information and Knowledge Management*, 1649–52. CIKM '09. New York, NY, USA: ACM. https://doi.org/10.1145/1645953.1646195.