The Journal of Open Source Software

Contextualized: Heterogeneous Modeling Toolbox

Caleb N. Ellington [1*¶], **Benjamin J. Lengerich** [2,3*¶], **Wesley Lo**[2,4], **Aaron Alvarez**[5], **Andrea Rubbi**[6], **Manolis Kellis**[2,3], **and Eric P. Xing**[1,7]

1 Carnegie Mellon University, USA **2** Massachusetts Institute of Technology, USA **3** Broad Institute of MIT and Harvard, USA **4** Worcester Polytechnic Institute, USA **5** University of Cincinnati, USA **6** Cambridge University, UK **7** Mohamed bin Zayed University of Artificial Intelligence, UAE **8** Petuum Inc., USA ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Fabian Scheipl

Reviewers:

- @holl-
- @pescap

Submitted: 12 February 2024
Published: unpublished

Contextualized
Heterogeneous Modeling Toolbox

Summary

Heterogeneous and context-dependent systems are common in real-world processes, such as those in biology, medicine, finance, and the social sciences. However, learning accurate and interpretable models of these heterogeneous systems remains an unsolved problem. Most statistical modeling approaches make strict assumptions about data homogeneity, leading to inaccurate models, while more flexible approaches are often too complex to interpret directly. Fundamentally, existing modeling tools force users to choose between accuracy and interpretability. Recent work on Contextualized Machine Learning (Lengerich et al., 2023) has introduced a new paradigm for modeling heterogeneous and context-dependent systems, which uses contextual metadata to generate sample-specific models, providing context-specific model-based insights and representing data heterogeneity with context-dependent model parameters.

Here, we present `Contextualized`, a SKLearn-style Python package for estimating and analyzing context-dependent models based on the Contextualized Machine Learning paradigm. `Contextualized` implements two reusable and extensible concepts: *a context encoder* which translates sample context or metadata into model parameters, and *sample-specific model* which is defined by the context-specific parameters. With the flexibility of context-dependent parameters, each context-specific model can be a simple model class, such as a linear or Gaussian model, providing direct model-based interpretability without sacrificing overall accuracy.

Statement of Need

`Contextualized` opens up new avenues for quantitative analysis of complex and heterogeneous data, and simplifies the process of transforming this data into results. In particular, `Contextualized`:

1. **Unifies Modeling Frameworks:** `Contextualized` unifies modeling approaches for both homogeneous and heterogeneous data, including population, sub-population, (latent) mixture, cluster-based, time-varying, and varying-coefficient models (Hastie & Tibshirani, 1993). Additionally, `Contextualized` naturally falls back to more traditional modeling frameworks when complex heterogeneity is not present. Not only is this convenient,

but it limits the number of modeling decisions and validation tests required by users, reducing the risk of misspecification and false discoveries (Lengerich et al., 2023).

2. **Models High-resolution Heterogeneity:** Contextualized models adapt to the context of each sample by using a context encoder, naturally accounting for high-dimensional, continuous, and fine-grained variation between samples (Ellington et al., 2023).
3. **Quantifies Heterogeneity in Data:** Context-specific models quantify the randomness and structure of the systems underlying each data point, and variation in context-specific model parameters quantifies the heterogeneity between data points (Al-Shedivat et al., 2018; Deuschel et al., 2023). `Contextualized` provides tools to analyze, test, and validate contextualized models, unlocking new studies of structured heterogeneity.
4. **Interpolates and Extrapolates to Unseen Contexts:** By using context encoders to translate between contextual information and model parameters, `Contextualized` learns meta-relationships between metadata and data. At test time, `Contextualized` can adapt to contexts which were never observed in the training data (Ellington et al., 2023).
5. **Analyzes Latent Processes:** By associating structured models with each sample, `Contextualized` enables analysis of samples with latent processes. These latent processes can be inferred from patterns in context-specific models, and can be used to identify latent subgroups, latent trajectories, and latent features that explain heterogeneity (Lengerich, Al-Shedivat, et al., 2022).
6. **Provides Direct Interpretability:** `Contextualized` estimates and analyzes context-specific statistical models. These statistical models are mathematically-constrained such that each parameter has specific meaning, permitting direct interpretation and immediate results (Lengerich, Nunnally, et al., 2022).
7. **Incorporates Multi-modal Data:** Context is a general and flexible concept, and context-encoders can be used to instill any type of contextual information into contextualized models, including images, text, tabular data, and more (Al-Shedivat et al., 2020; Lengerich et al., 2021; Lengerich, Al-Shedivat, et al., 2022; Stoica et al., 2020).
8. **Enables Modular Development:** The context encoder and sample-specific model within `Contextualized` are both highly adaptable; the context encoder can be replaced with any differentiable function, and any statistical model with a differentiable likelihood or log-likelihood can be contextualized and made sample-specific, benefiting from a rich ecosystem of statistical models and deep learning methods.

Usage

The `Contextualized` software is structured through three primary resources:

1. A simple plug-and-play interface to learn contextualized versions of popular model classes (e.g. classifiers, linear regression, graphical models, Gaussians).
2. A suite of context encoders to incorporate any modality of contextual data (e.g. continuous, categorical, images, text) and/or impose restrictions on context-dependent relationships (e.g. feature independence, interaction effects).
3. Intuitive analysis tools to understand, quantify, test, and visualize data with heterogeneous and context-dependent behavior. These tools focus on visualizing heterogeneity, automatic hypothesis testing, and feature selection for context-dependent and context-invariant features.

Installation instructions, tutorials, API reference, and open-source code are all available at contextualized.ml.

Acknowledgements

We are grateful for early user input from Juwayni Lucman, Alyssa Lee, and Jannik Deuschel.

References

Al-Shedivat, M., Dubey, A., & Xing, E. P. (2018). *Personalized Survival Prediction with Contextual Explanation Networks*. arXiv. https://doi.org/10.48550/arXiv.1801.09810

Al-Shedivat, M., Dubey, A., & Xing, E. P. (2020). *Contextual Explanation Networks*. arXiv. https://doi.org/10.48550/arXiv.1705.10301

Deuschel, J., Ellington, C. N., Lengerich, B. J., Luo, Y., Friederich, P., & Xing, E. P. (2023). *Contextualized Policy Recovery: Modeling and Interpreting Medical Decisions with Adaptive Imitation Learning*. arXiv. https://doi.org/10.48550/arXiv.2310.07918

Ellington, C. N., Lengerich, B. J., Watkins, T. B., Yang, J., Xiao, H., Kellis, M., & Xing, E. P. (2023). *Contextualized Networks Reveal Heterogeneous Transcriptomic Regulation in Tumors at Sample-Specific Resolution*. bioRxiv. https://doi.org/10.1101/2023.12.01.569658

Hastie, T., & Tibshirani, R. (1993). Varying-Coefficient Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *55*(4), 757–779. https://doi.org/10.1111/j.2517-6161.1993.tb01939.x

Lengerich, B. J., Al-Shedivat, M., Alavi, A., Williams, J., Labbaki, S., & Xing, E. P. (2022). *Discriminative Subtyping of Lung Cancers from Histopathology Images via Contextual Deep Learning*. medRxiv. https://doi.org/10.1101/2020.06.25.20140053

Lengerich, B. J., Ellington, C. N., Aragam, B., Xing, E. P., & Kellis, M. (2021). *NOTMAD: Estimating Bayesian Networks with Sample-Specific Structures and Parameters*. arXiv. https://doi.org/10.48550/arXiv.2111.01104

Lengerich, B. J., Ellington, C. N., Rubbi, A., Kellis, M., & Xing, E. P. (2023). *Contextualized Machine Learning*. arXiv. https://doi.org/10.48550/arXiv.2310.11340

Lengerich, B. J., Nunnally, M. E., Aphinyanaphongs, Y., Ellington, C., & Caruana, R. (2022). Automated Interpretable Discovery of Heterogeneous Treatment Effectiveness: A COVID-19 Case Study. *J. Biomed. Inform.*, 104086. https://doi.org/10.1016/j.jbi.2022.104086

Stoica, G., Stretcu, O., Platanios, E. A., Mitchell, T., & Póczos, B. (2020). Contextual Parameter Generation for Knowledge Graph Link Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(03), 3000–3008. https://doi.org/10.1609/aaai.v34i03.5693