CPH 200C: Computational Precision Health Cornerstone Thurs May 1, 2025

Problem Set 2: Interpretability

Instructor: Prof Irene Chen *Due date: Thurs May 15, 2025*

As machine learning models become more inscrutable, we are interested in methods to better understand why our models make the decisions they do — and as a result whether we can trust them. To address this problem, we typically apply either post-hoc explanations to trained models or restrict our models to specific classes that are deemed more easily understandable. The goal of this problem is to develop your ability to apply interpretability methods and communicate them to a technical audience. We will be leveraging the models we trained earlier in this class.

Instructions: This is not a group project and students will be graded individually. This project has two deliverables, which should be submitted on bCourses by 11:59pm PT Thurs May 15:

- **A written report**. The report should address the questions in Section 1 and Section 2.
- **A zip file with your code**. Please submit both your code and paper using bCourses. You will get feedback on both your paper and code via bCourses or email.

1 SHAP Values (20 pt)

We will be familiarizing ourselves with the use of SHAP values [LL17] for interpretability. See attached notebook for starter code loading the MIMIC-III dataset and training a distributed gradient boosting algorithm XGBoost on it. Although you can also run the script to create the MIMIC-III cohort, if you send Irene a screenshot of proof of your Physionet / MIMIC-III access, Irene will also send you a Dropbox download link of the MIMIC-III ICU mortality cohort that we are working with.

We are interested in predicting in-ICU mortality in MIMIC-III with data collected in the first 48 hours. You will create models to predict the in-ICU mortality using either: a) tabular data collected at admission and the tabular data collected during the first 48 hours or b) the clinical notes recorded within the first 48 hours.

1.1 SHAP values overview

What are SHAP values? What makes SHAP values different than linear coefficients for a logistic regression? What are the limitations on the model or data-type that can be used?

1.2 Tabular Data

Given a trained XGBoost algorithm, create a force plot that shows, for a specific data point, which tabular features are important. Over the whole model, which tabular features increase likelihood of in-ICU mortality – and which tabular features decrease the likelihood of in-ICU mortality?

1.3 Text Data

Given a TF-IDF feature vectorizer of the clinical notes, we are interested in evaluating how a support vector machine predicts in-ICU mortality. For the example sentence: "Pt is worsening with liver failure, will transfer to surgery", which notes increase the likelihood of in-ICU mortality? Which notes decrease the likelihood?

2 Other Interpretability Methods (20 pt)

Choose one interpretability method from the below list or another one that you would like to explore.

- Falling rule lists [WR15], including Github repo
- Influence functions [KL17], including Github repo
- Anchors [RSG18], including Github repo
- Actionable recourse [USL19], including Github repo
- Concept activation vectors [KWG+18], including Github repo
- Distillation to GAMs [TCHL18], including Github repo

2.1 Chosen interpretability method overview

Which method did you select? How does it work? What are the restrictions on models and datasets on which it can work?

2.2 Use on health data

Using the MIMIC-III data or another relevant health dataset, apply your interpretability method and present the results.

Interpretability methods are meant to serve towards a purpose of better understanding a model and dataset. As mentioned in the LIME paper [RSG16], potential options could include selecting the "best" classifier, removing out-of-distribution examples, improving trust in algorithms, and understanding why a model is making decisions. How would you design a user study to demonstrate the usefulness of this model? What metrics or other information would you collect from users in the study to determine if your method is improving on your research goal?

3 Feedback (1 pt)

How long did this problem set take you (in hours)? The 1 pt will be given for any response to this question, and the value will only be used for calibration of future problem sets.

References

[BKB17] Osbert Bastani, Carolyn Kim, and Hamsa Bastani. Interpreting blackbox models via model extraction. *arXiv preprint arXiv:1705.08504*, 2017.

[KL17] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[KWG+18] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[LL17] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[RSG18] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[TCHL18] Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. Distill-and-compare: Auditing black-box models using transparent model distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310, 2018.

[USL19] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 10–19, 2019.

[WR15] Fulton Wang and Cynthia Rudin. Falling rule lists. In *Artificial intelligence and statistics*, pages 1013–1022. PMLR, 2015.