Background: Recent advances in machine learning enable accurate classification for a variety of tasks. Data used to train these models often contains an excess number of features. Even though this works well on average; past work has identified that the presence of a *spurious correlation*, a connection between some features and the class label that seems causal but is not, pronounces this phenomenon [1]. As an example, consider the toy classification problem of predicting whether an image of a bird depicts a waterbird or landbird. It is reasonable to assume that in a training set, most waterbird images will have water backgrounds and most landbird images will have land backgrounds (I will refer to this as the waterbirds dataset). A model may latch onto the background and misclassify new data points from atypical groups (e.g. waterbirds with land background).

To tackle this problem, researchers have defined group distributionally robust optimization (gDRO), an objective to minimize the worst-case loss over pre-defined subsets of the training data [2]. In the landbird/waterbird example, one could train over 4 groups (one for each possible background – bird type pairing), so the model would learn to correctly classify even the worst-case group. Many practical datasets, however, do not have group information. For example, studies have shown that classification models for medical imaging often misclassify images from atypical groups, such as images of rare cancer subtypes, that were unknown during training [3].

My goal is to devise a theoretical framework to split a general dataset into groups for the gDRO objective given partial or no group information.

Related Work: One setting for learning groups for gDRO assumes groups labels are known for some subset of the training data. In [5], the authors define the BARACK framework, which uses the group labeled data to learn a model that predicts group labels for the rest of the data. Though the paper proves that using gDRO on these learned groups is within some error bound of using gDRO on the true groups, there is not much theory about how group accuracy affects the BARACK framework.

Another setting is when we have no group information. One approach, Just Train Twice (JTT) [6], first trains a model on the input data to predict the class label and splits it into two groups - one with the misclassified data points, and one with correctly classified data points. It then learns to predict the class label using gDRO on these two groups. JTT works well in the case where a single feature is responsible for the spurious correlation (such as the waterbirds dataset) but fails in more complex cases (e.g. multiple spurious features) and lacks a strong theory to support it [6]. Therefore, designing a theoretical framework for learning groups is still an open problem.

Research plan: Learning groups with no group information

I plan to use the following framework to learn groups without group information. First, I will learn a decision tree to predict the class label. Decision trees are structures that split data into varying groups in an attempt to classify them. Since spurious correlations seem easier to predict than class labels, I hypothesize that the nodes where decision trees split data may partition data

based on the spurious correlations. Thus, having learned a decision tree, for each ndoe of that tree, I will create a group that contains all data samples which satisfy the branch requirements to land at that node. Optimizing over all these groups with gDRO should prevent learning a model that relies on that correlation.

Because decision trees provide a useful theoretical framework to work with, it may be possible to analyze this group labeling algorithm theoretically. For example, one could assume that data comes from a distribution that includes a union of groups to show that given enough samples, the nodes of the decision tree will approximate the true groups. This may also rely on assumptions such as the learning algorithm used to learn the decision tree and the structure of the groups.

Broader Impacts: Improving gDRO and group labeling will have widespread impact. As discussed earlier, medical classification problems such as cancer detection would benefit from group labeling. Additionally, fairness theory suggests that splitting data by categories such as race and gender is useful in creating prediction models that do not discriminate on subpopulations with less training data. In practice however, these group labels can be expensive to obtain [4], making group learning necessary. In the domain of natural language interface, past work has shown that models often attach to spurious correlations associated with specific words when making predictions [7]. Trustworthy medical imaging, fairness, NLI, and other predictive models hold endless benefits for society. Theoretical guarantees on group labeling are essential in achieving this.

References

[1] S. Gururangan, et al. Annotation artifacts in natural language inference data. *In Association for Computational Linguistics (ACL)*, pp. 107–112, 2018.
[2] J. Duchi, et al. Statistics of robust optimization: a generalized empirical likelihood approach. *arXiv,* 2016.
[3] L. Oakden-Rayner, et al. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. *National Library of Medicine*, pp. 151-159, 2020.
[4] Improving Fairness without Demographic by Lagged Dynamic Grouping. *OpenReview*, 2022.
[5] N. Sohoni, et al. BARACK: Partially Supervised Group Robustness With Guarantees. *OpenReview,* 2022.
[6] E. Zheran Liu, B. Haghgoo, A. Chen, et al. Just Train Twice: Improving Group Robustness without Training Group Information. *Proceedings of Machine Learning Research*, 2021.
[7] R. McCoy, et al. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Interface. *ACL Anthology*, 2019.