A Learning Based Hypothesis Test for Harmful Covariate Shift

Tom Ginsberg · Zhongyuan Liang · Rahul G. Krishnan

UNIVERSITY OF TORONTO

Covariate Shift Detection with a Learning Model

Setup:

- A classifier f is trained on a dataset $\mathbf{P} = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \underset{\text{iid}}{\sim} \mathcal{P}$
- f is deployed on unlabeled samples $\mathbf{Q} = \{\tilde{x}_i\}_{i=1}^{m}$ where $\tilde{x}_i \underset{\text{iid}}{\sim} \mathcal{Q}$

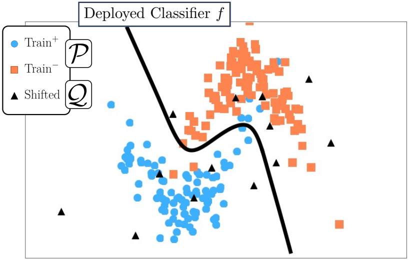
Covariate Shift Detection with a Learning Model

Setup:

- A classifier f is trained on a dataset $\mathbf{P} = \{(x_i, y_i)\}_{i=1}^{n}$ where $x_i \underset{\text{iid}}{\sim} \mathcal{P}$
- f is deployed on unlabeled samples $\mathbf{Q} = \{\tilde{x}_i\}_{i=1}^{m}$ where $\tilde{x}_i \underset{\text{iid}}{\sim} \mathcal{Q}$

Question:

- How can we leverage f to design a two-sample test for covariate shift between $\mathcal{P}$ and $\mathcal{Q}$ particularly when the observed number of test samples is small i.e. $|\mathbf{Q}| \ll |\mathbf{P}|$



Related Methods for Covariate Shift Detection

Dimensionality Reduction

- Autoencoder/PCA + Low Dimensional Two Sample Testing (Rabanser, Günnemann, and Z. C. Lipton, 2019)
- Black Box Shift Detection (Z. Lipton, Wang, and Smola, 2018)

Related Methods for Covariate Shift Detection

Dimensionality Reduction

- Autoencoder/PCA + Low Dimensional Two Sample Testing (Rabanser, Günnemann, and Z. C. Lipton, 2019)
- Black Box Shift Detection (Z. Lipton, Wang, and Smola, 2018)

Out-of-Distribution Detection / Uncertainty Estimation

- Deep Mahalanobis Score (Lee et al., 2018)
- Deep Ensembles (Ovadia et al., 2019)

Related Methods for Covariate Shift Detection

Dimensionality Reduction

- Autoencoder/PCA + Low Dimensional Two Sample Testing (Rabanser, Günnemann, and Z. C. Lipton, 2019)
- Black Box Shift Detection (Z. Lipton, Wang, and Smola, 2018)

Out-of-Distribution Detection / Uncertainty Estimation

- Deep Mahalanobis Score (Lee et al., 2018)
- Deep Ensembles (Ovadia et al., 2019)

High Dimensional Two-Sample Testing

- Classifier Two Sample Tests (Lopez-Paz and Oquab, 2017)
- Deep MMD (Liu et al., 2020)
- $\mathcal{H}$-Divergence (Zhao et al., 2022)

Related Methods for Covariate Shift Detection

Dimensionality Reduction

- Autoencoder/PCA + Low Dimensional Two Sample Testing (Rabanser, Günnemann, and Z. C. Lipton, 2019)
- Black Box Shift Detection (Z. Lipton, Wang, and Smola, 2018)

Out-of-Distribution Detection / Uncertainty Estimation

- Deep Mahalanobis Score (Lee et al., 2018)
- Deep Ensembles (Ovadia et al., 2019)

High Dimensional Two-Sample Testing

- Classifier Two Sample Tests (Lopez-Paz and Oquab, 2017)
- Deep MMD (Liu et al., 2020)
- $\mathcal{H}$-Divergence (Zhao et al., 2022)

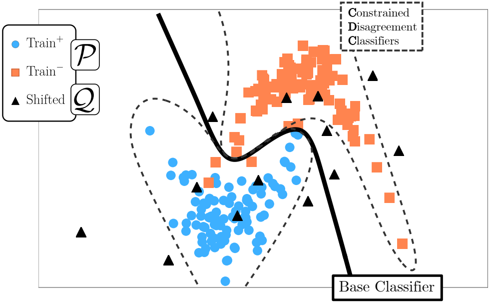
Shortcomings

$\rightarrow$ Not well suited to the small sample regime;
$\rightarrow$ Do not generalize for non Neural Network based models.

Constrained Disagreement

We build *constrained disagreement classifiers* (CDCs) to explicitly

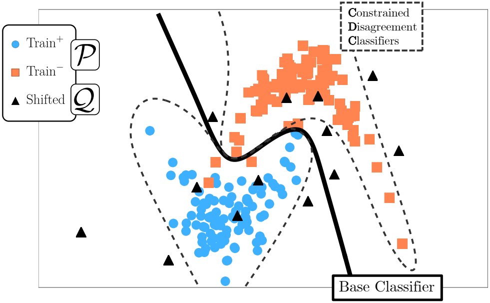
- Maximize out-of-distribution disagreement,
- while, constrained to behave similarly in the training domain



Constrained Disagreement

We build *constrained disagreement classifiers* (CDCs) to explicitly

- Maximize out-of-distribution disagreement,
- while, constrained to behave similarly in the training domain



We introduce the disagreement cross-entropy (DCE) as a score that encourages a classifier to disagree with a target label $y \in \{1, \ldots, N\}$

$$\mathrm{DCE}(\hat{y}, y) = \frac{1}{1-N} \sum_{i=1}^{N} \log(\hat{y}_i)\delta_{i \neq y}$$

The Detectron

A test for covariate shift using constrained disagreement

- We calibrate the expected disagreement rate of a CDC when trained to disagree on unseen data from the source distribution

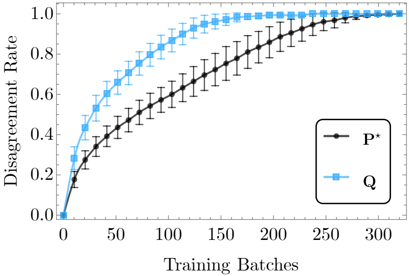


Figure: Training CDCs from a classifier f trained on CIFAR 10 to disagree on data unseen data from CIFAR 10 $\mathbf{P}^{\star}$ and near OOD data (CIFAR 10.1) $\mathbf{Q}$

The Detectron

A test for covariate shift using constrained disagreement

- We calibrate the expected disagreement rate of a CDC when trained to disagree on unseen data from the source distribution
- When training CDCs on out-of-distribution data **Q** we can reject the null hypothesis $\mathcal{P} = \mathcal{Q}$ if the disagreement rate is above the calibration

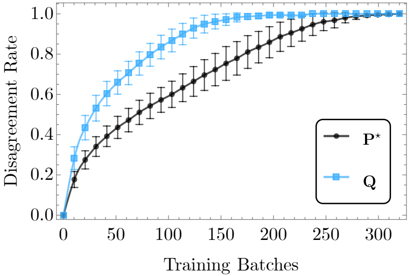


Figure: Training CDCs from a classifier f trained on CIFAR 10 to disagree on data unseen data from CIFAR 10 $\mathbf{P}^\star$ and near OOD data (CIFAR 10.1) $\mathbf{Q}$

Power of the Detectron Two-Sample Test

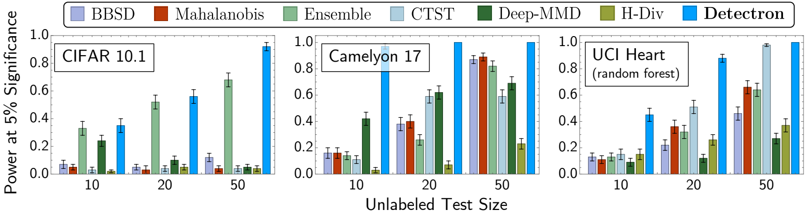


Figure: Detectron achieves compelling SOTA two sample testing results on several high-dimensional image and tabular datasets for extremely small sample sizes.

Concluding Remarks

- Our work presents a practical application for detecting covariate shifts that achieves SOTA performance on small sample sizes
- Our methodology works well on both neural networks and random forests

Concluding Remarks

- Our work presents a practical application for detecting covariate shifts that achieves SOTA performance on small sample sizes
- Our methodology works well on both neural networks and random forests

Future Directions

- Exploring the relationship between model complexity, generalization error and test power
- Improving the computational runtime
- Using constrained disagreement as a representation learning objective to correct for covariate shift
- Data efficient + learning-based methods for label and concept shift

Thank You

Rabanser, Stephan, Stephan Günnemann, and Zachary C. Lipton (2019). "Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*. Ed. by Hanna M. Wallach et al., pp. 1394–1406. URL: https://proceedings.neurips.cc/paper/2019/hash/846c260d715e5b854ffad5f70a516c88-Abstract.html.

Lipton, Zachary, Yu-Xiang Wang, and Alexander Smola (2018). "Detecting and correcting for label shift with black box predictors". In: *International conference on machine learning*. PMLR, pp. 3122–3130.

Lee, Kimin et al. (2018). "A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc.

Ovadia, Yaniv et al. (2019). "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc.

Lopez-Paz, David and Maxime Oquab (2017). "Revisiting Classifier Two-Sample Tests". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJkXfE5xx.

Liu, Feng et al. (2020). "Learning deep kernels for non-parametric two-sample tests". In: *International conference on machine learning*. PMLR, pp. 6316–6326.

Zhao, Shengjia et al. (2022). "Comparing Distributions by Measuring Differences that Affect Decision Making". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=KB5onONJIAU.