Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift

Stephan Rabanser rabanser@in.tum.de
Prof. Dr. Stephan Günnemann guennemann@in.tum.de
Prof. Zachary C. Lipton, PhD zlipton@cmu.edu

Technical University of Munich
Department of Informatics
Data Analytics and Machine Learning
http://daml.in.tum.de

January 8, 2020

Table of contents

Motivation & Overview

Motivation

- The reliable functioning of software depends crucially on tests.
- Despite their power, ML models are sensitive to shifts in the data distribution.
- ML pipelines rarely inspect incoming data for signs of distribution shift.
- Best practices for testing equivalence of the *source* distribution p and the *target* distribution q in real-life, high-dim. data settings have not yet been established.
- Existing solutions to addressing covariate shift

$$q(\boldsymbol{x}, y) = q(\boldsymbol{x})p(y|\boldsymbol{x})$$

or label shift

$$q(\boldsymbol{x}, y) = q(y)p(\boldsymbol{x}|y)$$

often rely on strict preconditions, producing wrong predictions if not met.

Shift Detection Overview

Faced with distribution shift, our goals are three-fold:

- detect when distribution shift occurs from as few examples as possible;
- characterize the shift (e.g. by identifying those samples from the test set that appear over-represented in the target data); and
- provide some guidance on whether the shift is harmful or not.

Methods

Our Framework

Given labeled data $(\boldsymbol{x}_1, y_1), ..., (\boldsymbol{x}_n, y_n) \sim p$ and unlabeled data $\boldsymbol{x}'_1, ..., \boldsymbol{x}'_m \sim q$, our task is to determine whether $p(\boldsymbol{x})$ equals $q(\boldsymbol{x}')$:

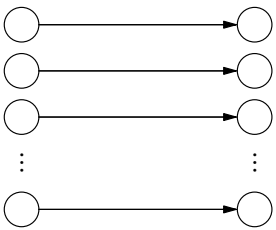
$$H_0 : p(\boldsymbol{x}) = q(\boldsymbol{x}') \qquad \text{vs} \qquad H_A : p(\boldsymbol{x}) \neq q(\boldsymbol{x}').$$

We explore the following design choices:

- what **representation** to run the test on;
- which **two-sample test** to run;
- when the representation is multidimensional; whether to run **a single multivariate test or multiple univariate two-sample tests**; and
- **how to combine** their results.

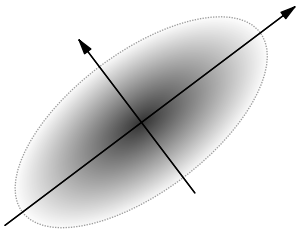
Dimensionality Reduction Techniques: NoRed & PCA

No Reduction (NoRed ○):



- To justify the use of any DR technique, our default baseline is to run tests on the original raw features.

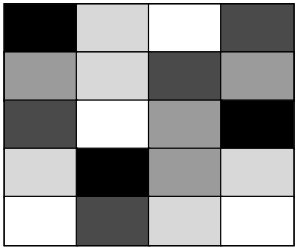
Principal Components Analysis (PCA ⬡):



- Find an optimal orthogonal transf. matrix such that points are linearly uncorrelated after transf.

Dimensionality Reduction Techniques: SRP & AE

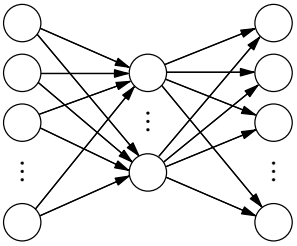
Sparse Random Projection (SRP ⬠):



$$R_{ij} = \begin{cases} +\sqrt{\frac{v}{K}} & \text{with prob. } \frac{1}{2v} \\ 0 & \text{with prob. } 1 - \frac{1}{v} \\ -\sqrt{\frac{v}{K}} & \text{with prob. } \frac{1}{2v} \end{cases}$$

with $v = \frac{1}{\sqrt{D}}$

Autoencoders (TAE ◇ and UAE □):

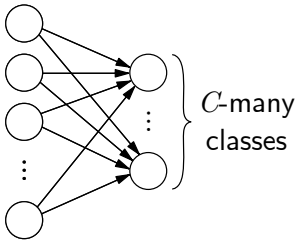


- Encoder $\phi : \mathcal{X} \to \mathcal{L}$
- Decoder $\psi : \mathcal{L} \to \mathcal{X}$

$$\phi, \psi = \arg\min_{\phi,\psi} \|\boldsymbol{X} - (\psi \circ \phi)\boldsymbol{X}\|^2$$

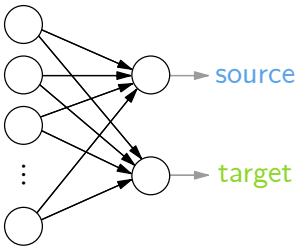
Dimensionality Reduction Techniques: BBSD & Classif

Label Classifiers (BBSDs ◁ and BBSDh ▷):



- Label classifier with softmax outputs (BBSDs ◁) or hard-thresholded predictions (BBSDh ▷).

Domain Classifier (Classif ×):

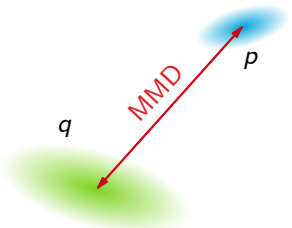


- Explicitly train a domain classifier to discriminate between data from source and target domains.

Statistical Hypothesis Testing: Maximum Mean Discrepancy (MMD)

- Popular kernel-based technique for multivariate two-sample testing.
- Distinguish two distrib. based on their mean embeddings $\boldsymbol{\mu}_p$ and $\boldsymbol{\mu}_q$ in a reproducing kernel Hilbert space $\mathcal{F}$:

$$\text{MMD}(\mathcal{F}, p, q) = ||\boldsymbol{\mu}_p - \boldsymbol{\mu}_q||^2_{\mathcal{F}}$$



- Empirical estimate:

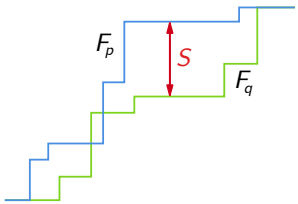
$$\begin{aligned}\text{MMD}^2 = &\frac{1}{m(m-1)} \sum_{i=1}^{m} \sum_{j \neq i}^{m} \kappa(\boldsymbol{x}_i, \boldsymbol{x}_j) \\ &+ \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \kappa(\boldsymbol{x}'_i, \boldsymbol{x}'_j) \\ &- \frac{2}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \kappa(\boldsymbol{x}_i, \boldsymbol{x}'_j)\end{aligned}$$

- Kernel: $\kappa(\boldsymbol{x}_1, \boldsymbol{x}_2) = e^{-\frac{1}{\sigma}\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^2}$
- Used with NoRed ○, PCA ⬡, SRP ⬠, TAE ◇, UAE □, and BBSDs ◁.

Statistical Hypothesis Testing: Kolmogorov-Smirnov + Bonferroni

- Test each of the K dimensions separately (instead of jointly) using the Kolmogorov-Smirnov (KS) test.
- Largest difference S of the cumulative density functions over all values $\boldsymbol{z}$:

$$S = \sup_{z} |F_p(\boldsymbol{z}) - F_q(\boldsymbol{z})|$$



- Multiple hypothesis testing: we must subsequently combine the p-values from the K-many test.
- Problem: We cannot make strong assumptions about the (in)dependence among the tests.
- Solution: Bonferroni correction:
  - Does not assume (in)dependence.
  - Bounds the family-wise error rate, i.e. it is a conservative aggregation.
  - Rejects H_0 if $p_{\text{min}} \leq \frac{\alpha}{K}$.
- Used with NoRed ○, PCA ⬡, SRP ⬠, TAE ◇, UAE □, and BBSDs ◁.

Statistical Hypothesis Testing: Chi-Squared Test

- Evaluate whether the freq. distr. of certain events observed in a sample is consistent with a particular theo. distr.
- Difference can be calculated as

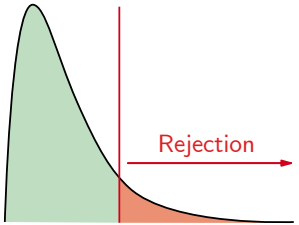
$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{C} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

with observed counts O_{ij} and expected counts $E_{ij} = N_{\text{sum}} p_{i\bullet} p_{\bullet j}$ with

- $p_{i\bullet} = \frac{n_{i\bullet}}{N_{\text{sum}}} = \sum_{j=1}^{c} \frac{n_{ij}}{N_{\text{sum}}}$ and
- $p_{\bullet j} = \frac{n_{\bullet j}}{N_{\text{sum}}} = \sum_{i=1}^{r} \frac{n_{ij}}{N_{\text{sum}}}$.

- Under H_0, $X^2 \sim \chi^2_{C-1}$.

Sample	Cat 1	$\cdots$	Cat C	$\sum$
p	n_{p1}	$\cdots$	n_{pC}	$n_{p\bullet}$
q	n_{q1}	$\cdots$	n_{qC}	$n_{q\bullet}$
$\sum$	$n_{\bullet 1}$	$\cdots$	$n_{\bullet C}$	N_{sum}



- Used with BBSDh ▷.

Statistical Hypothesis Testing: Binomial Test

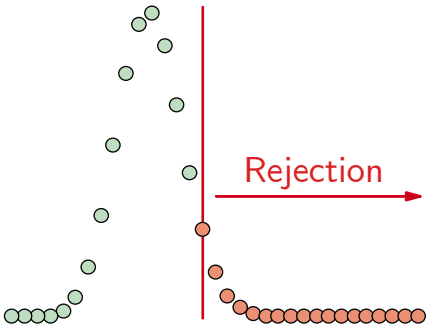
- Compare difference classifier accuracy (acc) on held-out data to random chance via a binomial test.

$$H_0 : \text{acc} = 0.5 \quad \text{vs} \quad H_A : \text{acc} > 0.5$$

- Under H_0, the acc follows a binomial distribution

$$\text{acc} \sim \text{Bin}(N_{\text{hold}}, 0.5)$$

where N_{hold} corresponds to the number of held-out samples.



- Used with Classif $\times$.

Obtaining Most Anomalous Samples

- Recall: our detection framework does not detect outliers but rather aims at capturing top-level shift dynamics.
- We can not decide whether any given sample is in- or out-of-distribution.
- But: we can harness domain assignments from the domain classifier.
- It is easy to identify the exemplars which the domain classifier was most confident in assigning to the target domain.

- Other shift detectors compare entire distributions against each other.
- Identification of samples which if removed would lead to a large increase in the overall p-value was not successful.

Determining the Malignancy of a Shift

- Distribution shifts can cause arbitrarily severe degradation in performance.
- In practice distributions shift constantly and often these changes are benign.
- Goal: distinguishing malignant shifts from benign shifts.
- Problem: although prediction quality can be assessed easily on source data, we are not able compute the target error directly without labels.

- Heuristic methods for approximating the target performance:
  - **Difference classifier assignments:** assess black-box model's accuracy on the labeled top anomalous samples (*implicit* shift characterization).
  - **Domain expert:** Get hints on the target accuracy by evaluating the classifier on held-out source data that has been *explicitly* perturbed by a function determined by a domain expert.

Experiments

Experimental Setup

- Core experiments: synthetic shifts on MNIST and CIFAR-10 image datasets.
- Autoencoders: convolutional architecture with 3 convolutional layers.
- BBSD and Classif: ResNet-18 architecture.
- Network training (TAE $\diamond$, BBSDs $\triangleleft$, BBSDh $\triangleright$, Classif $\times$): SGD with momentum in batches of 128 examples over 200 epochs with early stopping.
- Dimensionality reduction to $K = 32$ (PCA $\hexagon$, SRP $\pentagon$, UAE $\square$, and TAE $\diamond$), $C = 10$ (BBSDs $\triangleleft$), and 1 (BBSDh $\triangleright$ and Classif $\times$).
- Evaluate shift detection at a significance level of $\alpha = 0.05$.
- Shift detection performance is averaged over a total of 5 random splits.
- Randomly split the data into training, validation, and test sets and then apply a particular shift to the test set only.
- Evaluate the models with various amounts of samples from the test set $s \in \{10, 20, 50, 100, 200, 500, 1000, 10000\}$.

Shift Simulation

For each shift type (as appropriate) we explored three levels of shift intensity and various percentages of affected data $\delta \in \{0.1, 0.5, 1.0\}$.

- **Adversarial (adv)**: We turn a fraction δ of samples into adversarial samples via FGSM;
- **Knock-out (ko)**: We remove a fraction δ of samples from class 0, creating class imbalance;
- **Gaussian noise (gn)**: We corrupt covariates of a fraction δ of test set samples by Gaussian noise with standard deviation $\sigma \in \{1, 10, 100\}$ (denoted *s_gn*, *m_gn*, and *l_gn*);
- **Image (img)**: We also explore more natural shifts to images, modifying a fraction δ of images with combinations of random rotations $\{10, 40, 90\}$, (x, y)-axis-translation percentages $\{0.05, 0.2, 0.4\}$, as well as zoom-in percentages $\{0.1, 0.2, 0.4\}$ (denoted *s_img*, *m_img*, and *l_img*);
- **Image + knock-out (m_img+ko)**: We apply a fixed medium image shift with $\delta_1 = 0.5$ and a variable knock-out shift δ;

Shift Simulation (contd.)

- **Only-zero + image (oz+m_img)**: Here, we only include images from class 0 in combination with a variable medium image shift affecting only a fraction δ of the data;
- **Original splits**: We evaluate our detectors on the original source/target splits provided by the creators of MNIST, CIFAR-10, Fashion MNIST, and SVHN datasets (assumed to be i.i.d.);
- **Real shift datasets**:
  - Domain adaptation from MNIST (source) to USPS (target).
  - COIL-100 dataset where images between 0° and 175° are sampled by the source and images between 180° and 355° are sampled by the target distribution.

Dimensionality Reduction Methods Comparison

Table: Detection accuracy of different dimensionality reduction techniques across all simulated shifts on MNIST and CIFAR-10. **Green bold** entries indicate the best DR method at a given sample size, *red italic* the worst. Underlined entries indicate accuracy values > 0.5.

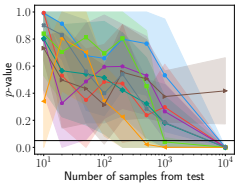
Test	DR	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univ. tests	NoRed	0.03	0.15	0.26	0.36	0.41	0.47	0.54	0.72
	PCA	0.11	0.15	0.30	0.36	0.41	0.46	0.54	0.63
	SRP	0.15	0.15	0.23	0.27	0.34	0.42	0.55	0.68
	UAE	0.12	0.16	0.27	0.33	0.41	0.49	0.56	0.77
	TAE	0.18	0.23	0.31	0.38	0.43	0.47	0.55	0.69
	BBSDs	**0.19**	**0.28**	**0.47**	**0.47**	**0.51**	**0.65**	**0.70**	**0.79**
χ^2	*BBSDh*	0.03	0.07	0.12	0.22	*0.22*	*0.40*	*0.46*	*0.57*
Bin	*Classif*	*0.01*	*0.03*	*0.11*	*0.21*	0.28	0.42	0.51	0.67
Multiv. tests	NoRed	0.14	*0.15*	*0.22*	*0.28*	0.32	*0.44*	0.55	–
	PCA	0.15	0.18	0.33	0.38	0.40	0.46	0.55	–
	SRP	*0.12*	0.18	0.23	0.31	*0.31*	*0.44*	0.54	–
	UAE	**0.20**	**0.27**	**0.40**	**0.43**	**0.45**	**0.53**	**0.61**	–
	TAE	0.18	0.26	0.37	0.38	0.45	0.52	0.59	–
	BBSDs	0.16	0.20	0.25	0.35	0.35	0.47	*0.50*	–

Shift Type Comparison

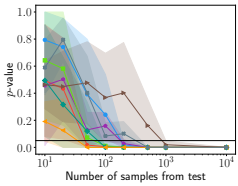
Table: Detection accuracy of different shifts on MNIST and CIFAR-10 using the best-performing DR technique (BBSDs). **Green bold** shifts are identified as harmless, *red italic* shifts as harmful.

Test	Shift	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univariate BBSDs	**s_gn**	0.00	0.00	0.03	0.03	0.07	0.10	0.10	0.10
	m_gn	0.00	0.00	0.10	0.13	0.13	0.13	0.23	0.37
	l_gn	0.17	0.27	0.53	0.63	0.67	0.83	0.87	1.00
	s_img	0.00	0.00	0.23	0.30	0.40	0.63	0.70	0.93
	m_img	0.30	0.37	0.60	0.67	0.70	0.80	0.90	1.00
	l_img	0.30	0.50	0.70	0.70	0.77	0.87	0.97	1.00
	adv	0.13	0.27	0.40	0.43	0.53	0.77	0.83	0.90
	ko	0.00	0.00	0.07	0.07	0.07	0.33	0.40	0.70
	m_img+ko	0.13	0.40	0.87	0.93	0.90	1.00	1.00	1.00
	oz+m_img	0.67	1.00	1.00	1.00	1.00	1.00	1.00	1.00

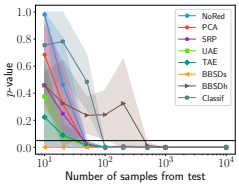
Individual Result: Medium Image Shift on MNIST



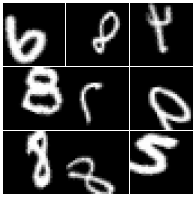
(a) Test w/ 10%.



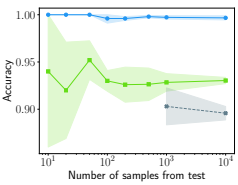
(b) Test w/ 50%.



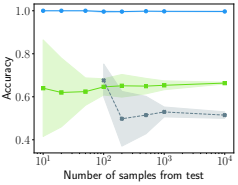
(c) Test w/ 100%.



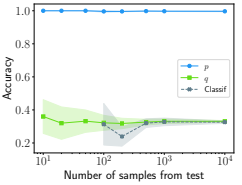
(d) Top different.



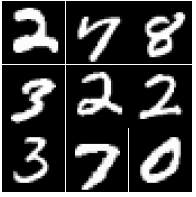
(e) Acc. w/ 10%.



(f) Acc. w/ 50%.

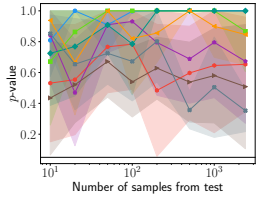


(g) Acc. w/ 100%.

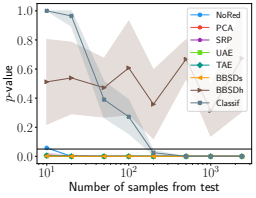


(h) Top similar.

Individual Result: Angle-Partitioning on COIL-100

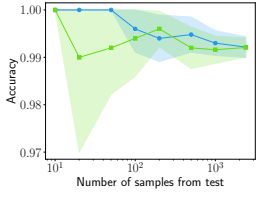


(a) Test random.

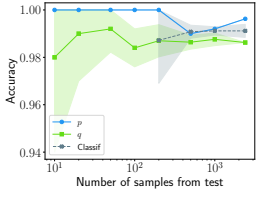


(b) Test partitioned.

(c) Top different.



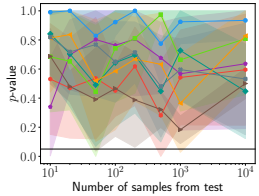
(d) Acc. random.



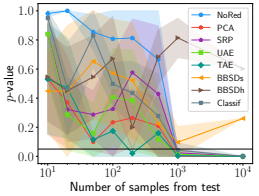
(e) Acc. partitioned.

(f) Top similar.

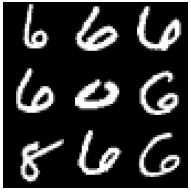
Individual Result: Original Split on MNIST



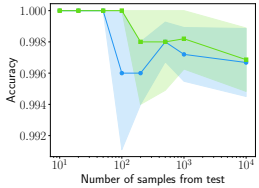
(a) Test random.



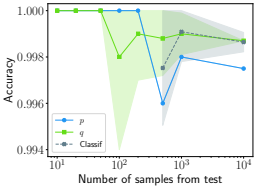
(b) Test original.



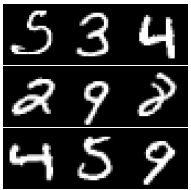
(c) Top different.



(d) Acc. random.



(e) Acc. original.



(f) Top similar.

Conclusion

Summary & Next Steps

Summary

- Black-box shift detection with soft predictions works well across many scenarios.
- Aggregated univariate tests performed separately on each latent dimension provide similar performance to multivariate two-sample tests, despite heavy correction.
- Harnessing predictions made by a domain-discriminating classifier enables characterization of the shift's nature and malignancy.

Potential future extensions

- Shift detection for online data by accounting for and exploiting the high degree of correlation between adjacent time steps.
- Apply our framework to other machine learning domains such as NLP or graphs.

Conference Submissions

DebugML @ ICLR 2019 (presented)

Presented at ICLR 2019 Debugging Machine Learning Models Workshop

FAILING LOUDLY: AN EMPIRICAL STUDY OF METHODS FOR DETECTING DATASET SHIFT

Stephan Rabanser[*], **Stephan Günnemann**
Technical University of Munich, Germany
{rabanser,guennemann}@in.tum.de

Zachary C. Lipton
Carnegie Mellon University, Pittsburgh, PA
zlipton@cmu.edu

ABSTRACT

We might hope that when faced with unexpected inputs, well-designed software systems would fire off warnings. Machine learning (ML) systems, however, which depend strongly on properties of their inputs (e.g. the i.i.d. assumption), tend to fail silently. This paper explores the problem of building ML systems that fail loudly, investigating methods for detecting dataset shift and identifying exemplars that most typify the shift. We focus on several datasets and various perturbations to both covariates and label distributions with varying magnitudes and fractions of data affected. Interestingly, we show that while classifier-based methods perform well in high-data settings, they perform poorly in low-data settings. Moreover, across the dataset shifts that we explore, a two-sample-testing-based approach, using pre-trained classifiers for dimensionality reduction performs best.

https://debug-ml-iclr2019.github.io/cameraready/DebugML-19_paper_20.pdf

NeurIPS 2019 (to be presented)

Failing Loudly: An Empirical Study of Methods for Detecting Dataset Shift

Stephan Rabanser[*], **Stephan Günnemann**
Technical University of Munich
{rabanser,guennemann}@in.tum.de

Zachary C. Lipton
Carnegie Mellon University
zlipton@cmu.edu

Abstract

We might hope that when faced with unexpected inputs, well-designed software systems would fire off warnings. Machine learning (ML) systems, however, which depend strongly on properties of their inputs (e.g. the i.i.d. assumption), tend to fail silently. This paper explores the problem of building ML systems that fail loudly, investigating methods for detecting dataset shift, identifying exemplars that most typify the shift, and quantifying shift malignancy. We focus on several datasets and various perturbations to both covariates and label distributions with varying magnitudes and fractions of data affected. Interestingly, we show that across the dataset shifts that we explore, a two-sample-testing-based approach, using pre-trained classifiers for dimensionality reduction, performs best. Moreover, we demonstrate that domain-discriminating approaches tend to be helpful for characterizing shifts qualitatively and determining if they are harmful.

https://arxiv.org/abs/1810.11953

Questions?

Thanks! :)

Backup

Multiple Hypothesis Testing Correction

Family-Wise Error Rate (FWER)

The most stringent control is given by procedures controlling the FWER, which limits the probability of making at least one false positive, formally

$$\text{FWER} = P(V \geq 1) < \alpha$$

where V is the total amount of false discoveries.

False Discovery Rate (FDR)

A less stringent but more powerful alternative to the FWER is the FDR, which limits the expected proportion of false positives, formally

$$\text{FDR} = \mathbb{E}\left[\frac{V}{M}\right] < \alpha$$

where M is the total amount of discoveries.

Shifts

Covariate Shift

$$[p(\boldsymbol{x}) \neq q(\boldsymbol{x}) \wedge p(y|\boldsymbol{x}) = q(y|\boldsymbol{x})] \Rightarrow p(y|\boldsymbol{x})p(\boldsymbol{x}) \neq q(y|\boldsymbol{x})q(\boldsymbol{x}) \Rightarrow p(\boldsymbol{x}, y) \neq q(\boldsymbol{x}, y)$$

Label Shift

$$[p(y) \neq q(y) \wedge p(\boldsymbol{x}|y) = q(\boldsymbol{x}|y)] \Rightarrow p(\boldsymbol{x}|y)p(y) \neq q(\boldsymbol{x}|y)q(y) \Rightarrow p(\boldsymbol{x}, y) \neq q(\boldsymbol{x}, y)$$

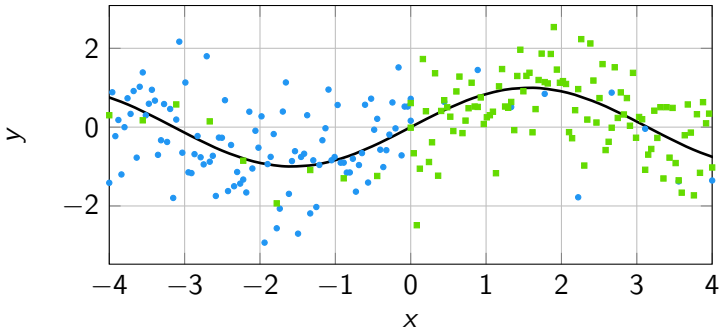
Concept Drift

$$[p(y|\boldsymbol{x}) \neq q(y|\boldsymbol{x}) \wedge p(\boldsymbol{x}) = q(\boldsymbol{x})] \Rightarrow p(y|\boldsymbol{x})p(\boldsymbol{x}) \neq q(y|\boldsymbol{x})q(\boldsymbol{x}) \Rightarrow p(\boldsymbol{x}, y) \neq q(\boldsymbol{x}, y)$$

$$[p(\boldsymbol{x}|y) \neq q(\boldsymbol{x}|y) \wedge p(y) = q(y)] \Rightarrow p(\boldsymbol{x}|y)p(y) \neq q(\boldsymbol{x}|y)q(y) \Rightarrow p(\boldsymbol{x}, y) \neq q(\boldsymbol{x}, y)$$

Covariate Shift

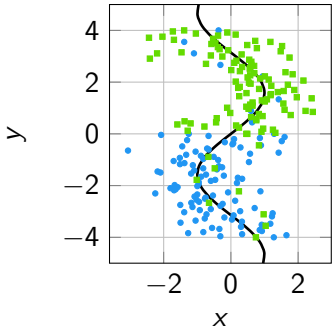
(a) Covariate shift causal graphical model.



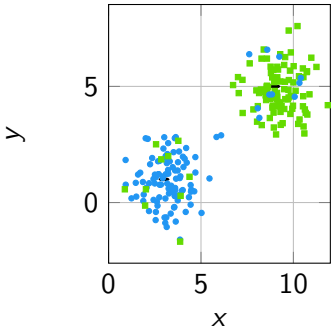
(b) Covariate shift example.

Label Shift

(a) Label shift causal graphical model.



(b) Regression example.



(c) Classification example.

Shift Intensity Comparison

Table: Detection accuracy for small, medium, and large simulated shifts and low (10%), medium (50%), and high (100%) percentages of perturbed target samples on MNIST and CIFAR-10. Reported accuracy values are results of the best DR technique (univariate: BBSDs, multivariate: average of UAE and TAE). Underlined entries indicate accuracy values > 0.5.

Test	Intensity	Number of samples from test							
		10	20	50	100	200	500	1,000	10,000
Univariate	Small	0.00	0.00	0.14	0.14	0.18	0.36	0.40	0.54
	Medium	0.14	0.21	0.39	0.38	0.42	0.57	0.66	0.76
	Large	0.32	0.54	0.78	0.82	0.83	0.92	0.96	1.00
	10%	0.11	0.15	0.24	0.25	0.28	0.44	0.54	0.66
	50%	0.14	0.28	0.52	0.53	0.60	0.68	0.72	0.85
	100%	0.26	0.41	0.61	0.64	0.70	0.82	0.84	0.86
Multivariate	Small	0.11	0.11	0.12	0.14	0.20	0.23	0.33	–
	Medium	0.11	0.19	0.23	0.27	0.32	0.42	0.44	–
	Large	0.34	0.45	0.57	0.68	0.72	0.82	0.93	–
	10%	0.12	0.13	0.21	0.26	0.27	0.31	0.44	–
	50%	0.19	0.27	0.41	0.41	0.47	0.57	0.60	–
	100%	0.29	0.41	0.44	0.53	0.60	0.70	0.78	–

MNIST Difference Plot

- The original splits from the MNIST dataset appear to exhibit a dataset shift.
- We observed that the top anomalous samples depicted the digit 6.
- This particular shift does not look significant to the human eye and is also declared harmless by our malignancy detector.