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

Stephan Rabanser\textsuperscript{1} Stephan G"unnemann\textsuperscript{2} Zachary C. Lipton\textsuperscript{3}
\textsuperscript{1}AWS AI Labs \textsuperscript{2}Technical University of Munich \textsuperscript{3}Carnegie Mellon University

**Motivation**

- The reliable functioning of software depends crucially on tests (unit tests, input validation).
- Despite their power, many machine learning models are sensitive to shifts in the data distribution.
- In practice, ML pipelines rarely inspect incoming data for any signs of distributional shift.
- Best practices for detecting shift in high-dimensional and real-life data have not yet been established.
- Existing solutions to addressing isolated shifts like covariate shift or label shift crucially on tests (unit tests, input validation).

**Our Framework**

Given labeled data \( \{(x_i, y_i), \ldots, (x_n, y_n)\} \sim p \) and unlabeled data \( \{x'_i, \ldots, x'_m\} \sim q \), our task is to determine whether \( p(x) \) equals \( q(x) \):
\[
H_0 : p(x) = q(x) \quad \text{vs} \quad H_A : p(x) \neq q(x).
\]

We explore the following design considerations:
- what representation to run the test on;
- which statistical 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**

- **(a) No Reduction (NoRed)**
- **(b) Principal Components (PCA)**
- **(c) Random Projection (SRP)**
- **(d) Autoencoder (TAE)**
- **(e) Classifier (BBSDs)**
- **(f) Domain Classifier (Classif.)**

**Statistical Two-Sample Testing**

- **(a) Max Mean Discrep. (MMD)**
- **(b) Kolmogorov-Smirnov Test (KS)**
- **(c) Chi-Squared Test \( \chi^2 \)**
- **(d) Binomial Test (Bin)**

**Key Observations**

- Multiple univariate tests and multivariate kernel tests offer comparable detection performance.
- **BBSDs (univariate)** and **UAE (multivariate)** are the best-performing shift detectors, respectively.
- Top different samples from **Classif.** are helpful in characterizing a shift’s nature and malignancy.
- MNIST original split is not i.i.d., but harmless.

**Shift Detection Pipeline Overview**

**Key Results**

- **(a) Detection accuracy of different dimensionality reduction techniques across all simulated shifts on MNIST and CIFAR-10.**
- **(b) Detection accuracy of different shifts on MNIST and CIFAR-10 using the best DR techniques. Green bold shifts are identified as harmless, red italic shifts as harmful.**

---

Stephan Günnemann 1
Carnegie Mellon University