Leveraging Common Structure to Improve Prediction Across Related Datasets

Matt Barnes, Nick Gisolfi, Madalina Fiterau, Artur Dubrawski

Auton Lab | Carnegie Mellon University

Introduction and Problem Formulation

Contradictions between data sets are common and their causes are widespread. Contradictions negatively impact model performance.

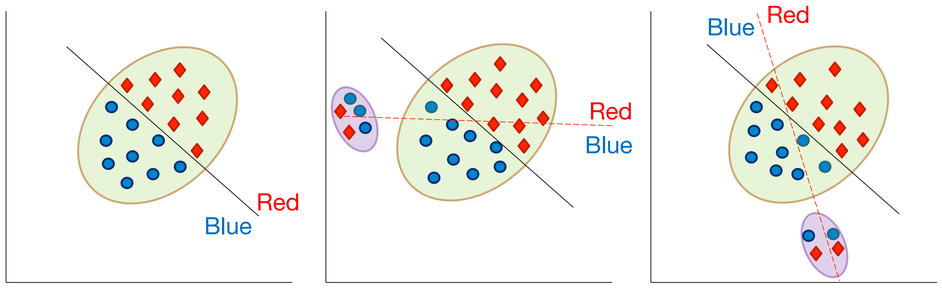
A majority of state of the art research in ML focuses on building better models, however, less work is focused on building better training sets.

Intuitively, more accurate models may be obtained by using data without contradicting labels or structured outliers. We present a simple approach to remove spurious samples by simultaneously analyzing multiple datasets.

Concept: Common and Spurious Distributions

Consider the task of training a binary classifier over multiple data sets:

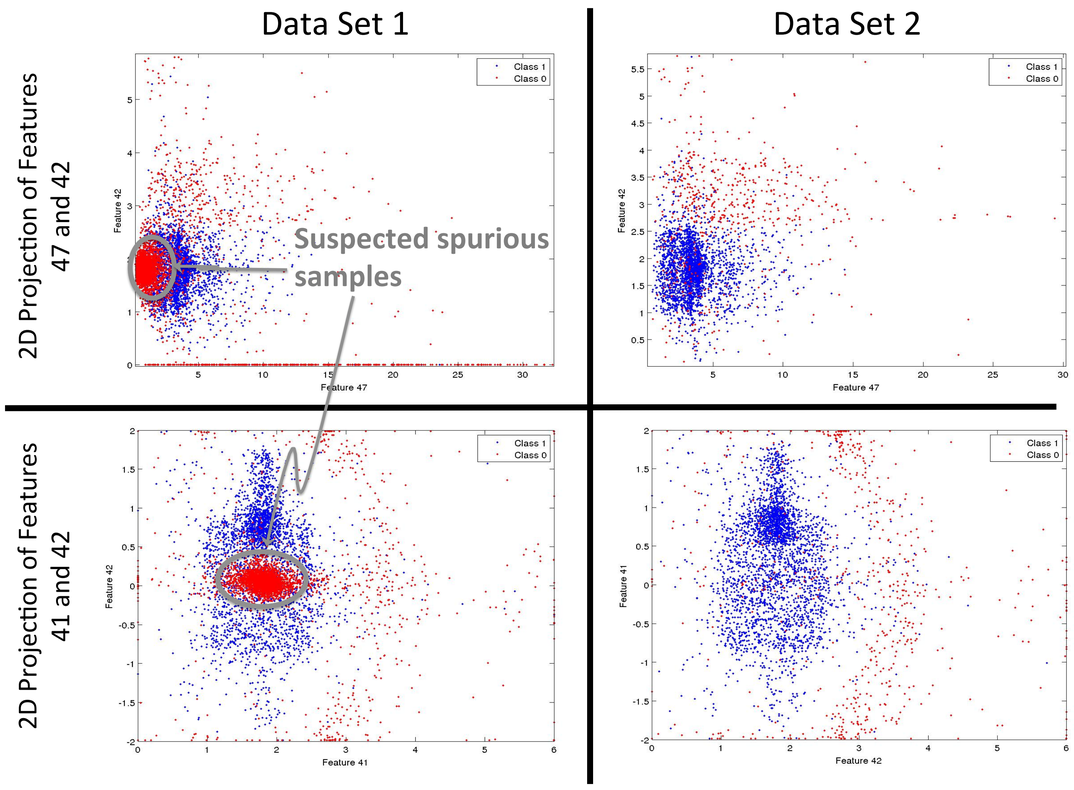
The common distribution is consistent throughout all data sets.
Spurious distributions appear in some data sets, but not all sets.



Learning on the common structure | Learning on a dataset A with a spurious distribution | Learning on a dataset B with a spurious distribution

Spurious Distributions in Real World Data

Low dimensional projections of data from separate sources



Datasets 1 and 2 are a good example of what we consider to be data with spurious distributions. The dense Class 0 (red) population is not present in datasets 2-5 (not all shown).

Hypothesis: Removing spurious distributions will improve a model's classification accuracy on future datasets

Methods

First, we defined global divergence as the sum of all pairwise divergence estimates:

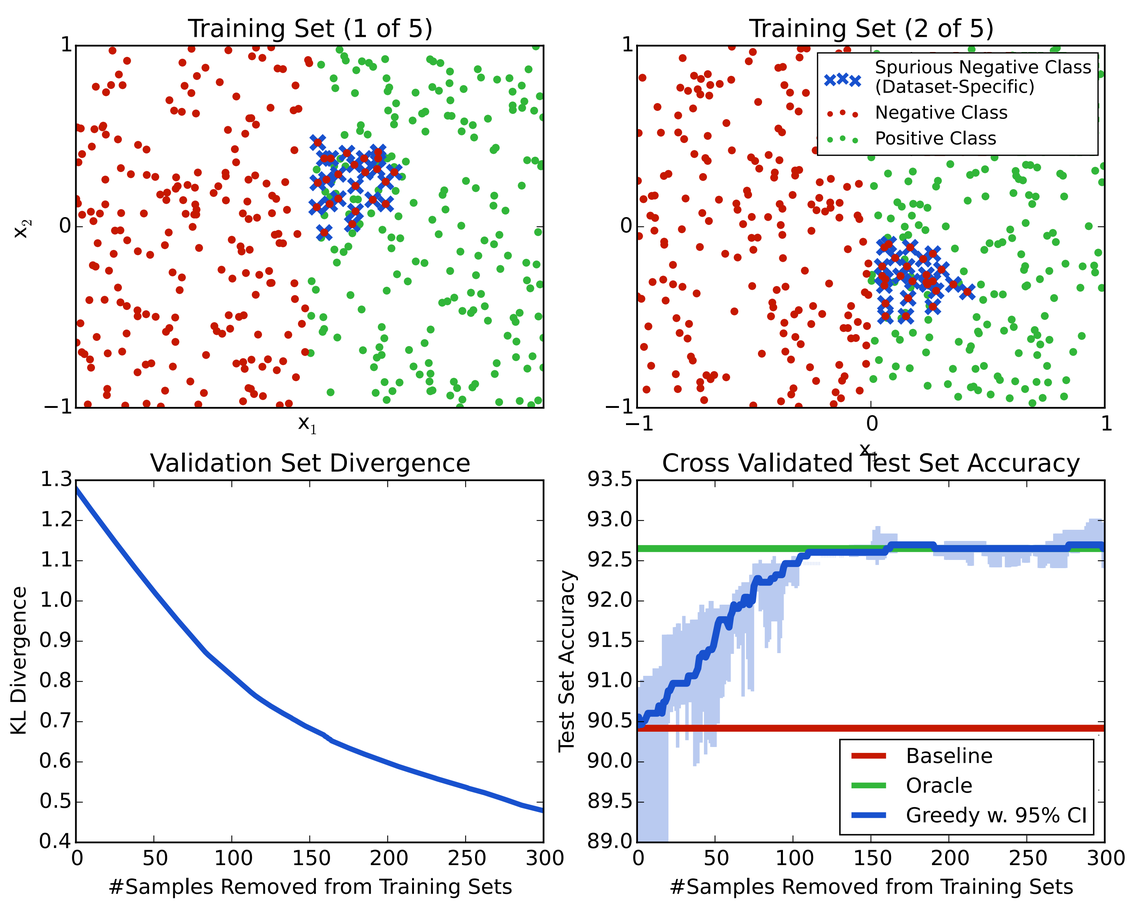
$$D_{global} = \sum_{i=1}^{N} \sum_{j=1}^{N} D\left(X_i \,\|\, X_j\right)$$

where $D(X_i \,||\, X_j)$ is some density estimator, usually Renyi. Then, we greedily removed samples which most reduced this global divergence. For high dimensional data, we used informative low dimensional projections because of dimensionality limitations with some density estimators (e.g. Kernel Density Estimate performance declined beyond approximately six dimensions).

As spurious samples are removed from the training set the algorithm learns the common structure. Removing further samples yield diminishing returns of divergence reduction. The process is terminated when divergence is below an acceptable level or when the change point is detected. For the experiments shown, we removed many more samples to demonstrate the plateau in testing accuracy.

Based on the context of the real world data and prior experiments, model performance was measured with an SVM or random forest.

Results



Greedy sample removal using HDE removes spurious samples in contradicting datasets (top left and right). A change-point in the validation set KL-divergence exists between spurious and default samples (bottom left). Cross validated tests on untouched datasets with their own unique spurious distributions showed a clear increase in accuracy (bottom right). The baseline system was trained on the original data, the oracle system was trained on only the common structure, and our greedy system learned the common structure in the training data.