Tracking the risk of a deployed model and detecting harmful distribution shifts

Aleksandr Podkopaev, Aaditya Ramdas

Carnegie Mellon University

Carnegie Mellon University

Setup

Deployed machine learning models inevitably encounter changes in distribution. It often may make sense to ignore benign shifts, under which the performance of a model does not degrade substantially, making interventions, such as model retraining, unnecessary. We differentiate between malignant and benign shifts by measuring changes in a user-specified metric, like accuracy or calibration.

We design nonparametric sequential hypothesis tests that (a) provably control the false alarm rate despite the multiple testing issues caused by continuous monitoring and (b) do not constrain the form of allowed shifts.

Detection as a sequential testing problem

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the covariate and label spaces respectively. Let $\ell(\cdot, \cdot)$ be the loss function chosen to be monitored, with $R(f) := \mathbb{E}\left[\ell(f(X), Y)\right]$ denoting the risk of a predictor $f : \mathcal{X} \to \mathcal{Y}$. We also consider running risk:

$$R^{(t)}(f) = \frac{1}{t}\sum_{i=1}^{t} \mathbb{E}\left[\ell\left(f(X_i'), Y_i'\right)\right], \quad t \geqslant 1,$$

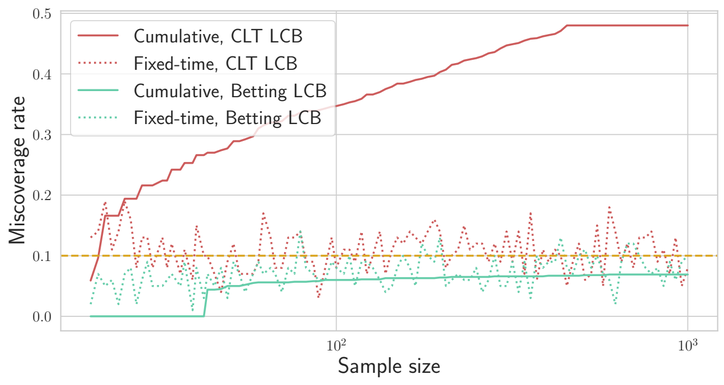
where the expected value is taken with respect to the joint distribution of (X_i', Y_i'), possibly different for each test point i. We aim to test:

$$H_0 : \; R_T^{(t)}(f) \leqslant R_S(f) + \varepsilon_{\text{tol}}, \; \forall t \geqslant 1,$$
$$H_1 : \; \exists t^\star \geqslant 1 : R_T^{(t^\star)}(f) > R_S(f) + \varepsilon_{\text{tol}},$$

where ε_{tol} is a tolerance level, $R_S(f)$ and $R_T^{(t)}(f)$ denote the source and target risks.

Suppose that one observes a sequence of data points $Z_1, Z_2, \ldots$. At each time point t, a sequential test takes the first t elements of this sequence and output either a 0 (continue) or 1 (reject the null and stop). Formally, a level-δ sequential test Φ defined as a mapping $\bigcup_{n=1}^{\infty} \mathcal{Z}^n \to \{0, 1\}$ must satisfy: $\mathbb{P}_{H_0}(\exists t \geqslant 1 : \Phi(Z_1, ..., Z_t) = 1) \leqslant \delta$, that is, if the null H_0 is true, then the probability that the test ever outputs a 1 and stops (false alarm) is at most δ.

Traditional (fixed-time) testing procedures are not valid under sequential settings and require corrections for multiple testing. Instead, we utilize confidence sequences which allow for continuous monitoring of model performance.



However, naive corrections for multiple testing do not take advantage of the dependence between the tests, and thus lead to losses of power of the resulting procedure.

Framework overview

Risk on the source is usually assessed through a labeled holdout sample of a fixed size: $\{(X_i, Y_i)\}_{i=1}^{n_S}$. Classic concentration results give an upper bound $\widehat{U}_S(f)$ on the risk:

$$\mathbb{P}\left(R_S(f) \leqslant \widehat{U}_S(f)\right) \geqslant 1 - \delta_S.$$

Target risk has to be re-estimated as losses on test instances are observed. Time-uniform confidence sequences retain validity under adaptive settings and give a time-uniform lower bounds $\widehat{L}_T^{(t)}(f)$, $t = 1, 2, \ldots$ on the risk:

$$\mathbb{P}\left(\exists t \geqslant 1 : R_T^{(t)}(f) < \widehat{L}_T^{(t)}(f)\right) \leqslant \delta_T.$$

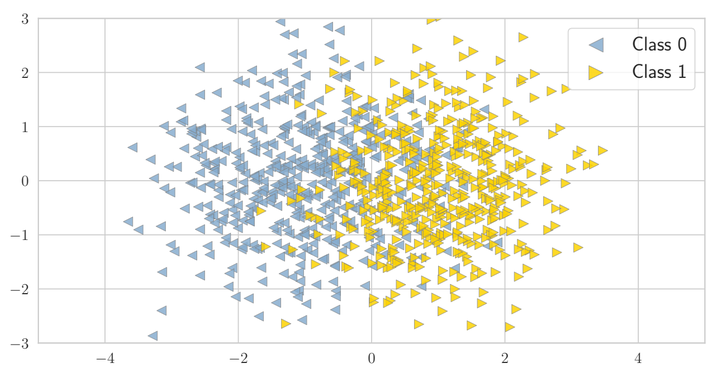
Procedure: Compute the upper confidence bound on the source risk $\widehat{U}_S(f)$ at level δ_S. For each time point $t = 1, 2, \ldots$

1. Compute the lower confidence bound on the target risk $\widehat{L}_T^{(t)}(f)$ at level δ_T.
2. Compute:
$$\Phi(Z_1, ..., Z_t) = \mathbb{1}\left\{\widehat{L}_T^{(t)}(f) > \widehat{U}_S(f) + \varepsilon_{\text{tol}}\right\}$$
3. If $\Phi(Z_1, ..., Z_t) = 1$, reject H_0 and fire off a warning.

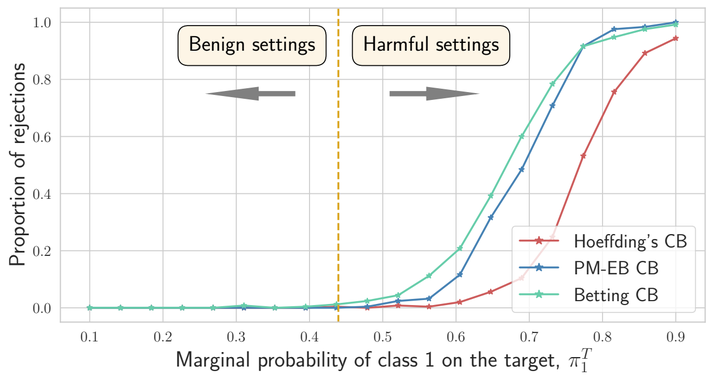
Proposition 1. *The proposed procedure controls the type I error uniformly over time:*

$$\mathbb{P}_{H_0}\left(\exists t \geqslant 1 : \Phi(Z_1, ..., Z_t) = 1\right) \leqslant \delta_S + \delta_T.$$

Consider a binary classification problem where data are sampled from a Gaussian mixture: $X \mid Y = y \sim \mathcal{N}(\mu_y, I_2)$, and by design, the classes overlap. For a fixed marginal probability of class 1 on the source ($\pi_1^S = 0.25$), we use the corresponding Bayes-optimal predictor. We induce label shift on the target domain, which malignancy is fully determined by the value of π_1^T.



Variance-adaptive bounds are much tighter when the individual losses $\ell(f(X_i), Y_i)$ have low variance. As a result, harmful shifts are detected much earlier, while the false alarm rate is still controlled at any prespecified level.



Simulations

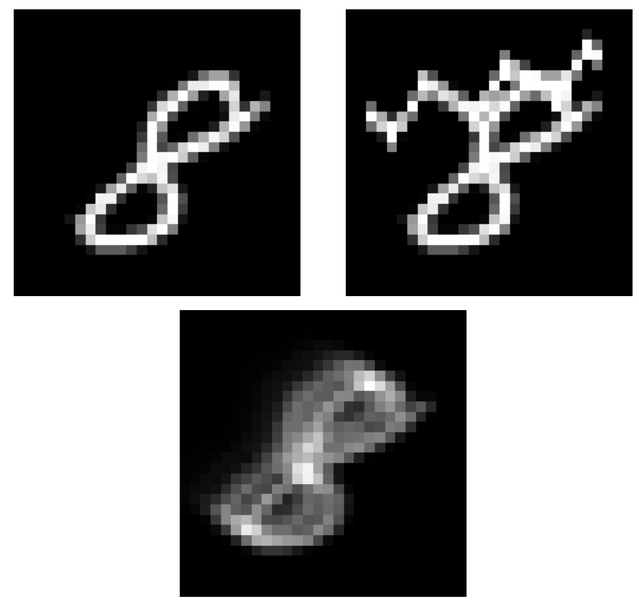


Figure 1: *Examples of corrupted CIFAR-10 images.*

Using a shallow CNN trained on clean MNIST data, we test whether the misclassification risk increases by 10% when clean and corrupted images are passed to the network. Not all corruptions are found to represent harmful shifts.

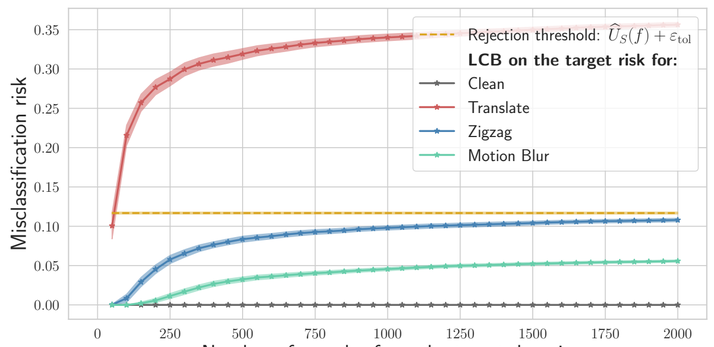


Figure 2: *Examples of corrupted CIFAR-10 images.*

We build a wrapper around a ResNet-32 model which outputs a set of candidate labels as a prediction. Under the i.i.d. assumption, it is guaranteed to have low miscoverage risk (0.1) with high probability (at least 95%). We test whether coverage drops by 5% when clean and corrupted images are passed to the model. Only the most intense level of fog is found to be consistently harmful to coverage.