Review for
"Black box tests for algorithmic stability"

Jiaxin Hu

1 High level summary

Motivation:

Many modern learning algorithms are too complex to analytically study the underlying mechanism and thus to evaluate the algorithm stability. The data distribution is also unknown in practice. Hence, we need an algorithm-free distribution-free method to evaluate or test the stability of the so-called "black box" algorithms via empirical performance only.

Take-away:

1. Propose a rigours algorithm stability definition depended on the sample distribution and size;
2. Formulate the stability evaluation for black box algorithms into a hypothesis testing problem and propose an algorithm-free and distribution-free test;
3. Provide the upper bound of the power for all valid stability tests with controlled type-I error;
4. Show that the naive Binomial test which evenly splits the data sets achieves the optimal power.

Limitations:

1. The theoretical analysis only applies for the *symmetric* algorithms; i.e., the prediction result is independent with the order of the data;
2. The analysis relies on the i.i.d. setting of the available data;
3. The analysis focuses on the data with domain $\mathbb{R}^d$ for some constant $d \geq 1$, and thus the analysis may not be optimal for discrete data (e.g. Bernoulli data);
4. This work assumes the labelled (training) data and unlabelled (testing) data share the same predictor distribution, which may not be true in practice.

2 Brief summary for method

Notation

1. Let $\{(X_i, Y_i)\}_{i\in[n]} \in \bigcup_{n\geq 1}(\mathbb{R}^d \times \mathbb{R})^n$ denote the data set to train the algorithm and n is the number of sample size that we want to test in stability; assume $(X_i, Y_i) \sim_{i.i.d.} P$, where P is the unknown data distribution;

2. Let $\mathcal{D}_l = \{(X_i, Y_i)\}_{i\in[N_l]}$ and $\mathcal{D}_u = \{X_i\}_{i\in N_u}$ denote the labelled and unlabelled data, respectively; let $\kappa = \min N_l/n, (N_l + N_u)/(n + 1)$ denote the number of copies of independent datasets constructed by $\mathcal{D}_u$ and $\mathcal{D}_l$.

3. Let $\mathcal{A} : \{(X_i, Y_i)\}_{i\in[n]} \mapsto \{\hat{\mu} : \mathbb{R}^d \mapsto \mathbb{R}\}$ denote the black box algorithm and $\hat{\mu}$ is the trained regression model. We assume $\mathcal{A}$ is symmetric; i.e., $\mathcal{A}(\{(X_i, Y_i)\}_{i\in[n]}) = \mathcal{A}(\{(X_{\pi(i)}, Y_{\pi(i)})\}_{i\in[n]})$, where π is a permutation on $[n]$.

- For simplicity, we ignore the discussion about the randomized algorithm (e.g. stochastic gradient descent) in this note; in general, we need to include an extra "random seed" parameter in the algorithm function.

- For simplicity, we focus on the case with adequate available data and κ is an integer; i.e., $\kappa \in \mathbb{Z}_+$. See original paper for the case and notation when κ is not an integer or $\kappa < 1$.

Definition 1 (Algorithm stability)**.** Let $\mathcal{A}$ be a symmetric algorithm. Let fixed $\epsilon \geq 0$ and $\delta \in [0, 1]$. We say the triplet $(\mathcal{A}, P, n)$ is (ϵ, δ)-stable if

$$\mathbb{P}(|\hat{\mu}_n(X_{n+1}) - \hat{\mu}_{n-1}(X_{n+1})| > \epsilon) \leq \delta,$$

where $\hat{\mu}_n$ and $\hat{\mu}_{n-1}$ are the fitted models trained by datasets $\{(X_i, Y_i)\}_{i\in[n]}$ and $\{(X_i, Y_i)\}_{i\in[n-1]}$, respectively, and $(X_i, Y_i) \sim_{i.i.d.} P$.

Binomial test

We consider the hypothesis testing:

$$H_0 : \text{the triplet } (\mathcal{A}, P, n) \text{ is not } (\epsilon, \delta)\text{-stable} \quad vs \quad H_1 : \text{the triplet } (\mathcal{A}, P, n) \text{ is } (\epsilon, \delta)\text{-stable.}$$

We summarize the naive Binomial test here; see Section 2 for rigorous definition of black box testing and other examples.

1. Construct κ independent datasets $\mathcal{D}^{(k)}, k \in [\kappa]$ with n labelled data and 1 unlabelled data; i.e., $\mathcal{D}^{(k)} = \{(X_i, Y_i)\}_{i=(k-1)n+1}^{kn} \cup \mathcal{D}_u^{(k)} = \{X_{\kappa n+k}\}$;

2. Obtain the fitted models $\hat{\mu}_n^{(k)} = \mathcal{A}(\{(X_i, Y_i)\}_{i=(k-1)n+1}^{kn})$ and $\hat{\mu}_{n-1}^{(k)} = \mathcal{A}(\{(X_i, Y_i)\}_{i=(k-1)n+1}^{kn-1})$. Calculate the difference

$$\Delta^{(k)} = |\hat{\mu}_n^{(k)}(X_{\kappa n+k}) - \hat{\mu}_{n-1}^{(k)}(X_{\kappa n+k})|.$$

3. Count the number of large $\Delta^{(k)}$

$$B = \sum_{k\in[\kappa]} \mathbb{1}\{\Delta^{(k)} > \epsilon\}.$$

Next, we summarize the theoretical property of the proposed Binomial test. Let δ_ϵ^* denote the true probability

$$\delta_\epsilon^* = \mathbb{P}(|\hat{\mu}_n(X_{n+1}) - \hat{\mu}_{n-1}(X_{n+1})| > \epsilon).$$

We reject the null hypothesis when B is smaller than the critical value k^* depends on κ and δ. We consider the test $\hat{T}_{\epsilon,\delta} = \mathbb{1}B < k^* + a^*\mathbb{1}\{B = k^*\}$, where a^* is a parameter depended on κ, δ and confidence level α to control the type-I error.

Theorem 2.1 (Property of Binomial test)**.** *Fix ϵ, $\delta \in [0, 1)$ and the confidence level $\alpha \in (0, 1)$. The Binomial test $\hat{T}_{\epsilon,\delta}$ has type-I error α; i.e.,*

$$\mathbb{P}(\hat{T}_{\epsilon,\delta} = 1 | (\mathcal{A}, P, n) \text{ is not } (\epsilon, \delta)\text{-stable }) \leq \alpha.$$

If $\delta^* = 0$ or $\delta \leq 1 - \alpha^{1/\kappa}$, we have power

$$\mathbb{P}(\hat{T}_{\epsilon,\delta} = 1 | (\mathcal{A}, P, n) \text{ is } (\epsilon, \delta)\text{-stable }) = \left\{\alpha \left(\frac{1 - \delta_\epsilon^*}{1 - \delta}\right)^\kappa\right\} \bigwedge 1.$$

Theorem 2.2 (Power of all black box test)**.** *Fix ϵ, $\delta \in [0, 1)$ and the confidence level $\alpha \in (0, 1)$. For all black box test $\hat{T}_{\epsilon,\delta}$ with type-I error smaller than 1, the power of the test $\hat{T}_{\epsilon,\delta}$ satisfies*

$$\mathbb{P}(\hat{T}_{\epsilon,\delta} = 1 | (\mathcal{A}, P, n) \text{ is } (\epsilon, \delta)\text{-stable }) = \left\{\alpha \left(\frac{1 - \delta_\epsilon^*}{1 - \delta}\right)^\kappa\right\} \bigwedge 1.$$

Compared with two theorems, we conclude that the naive Binomial test achieves optimal power. This is surprising since more sophisticated resampling procedures (compared with the random partition in Binomial test) and more runs of $\mathcal{A}$ (Binomial test only run two times for each dataset to calculate $\hat{\mu}_n$ and $\hat{\mu}_{n-1}$) do not help to improve the power of the test.

3 Detailed Questions

1. (Optimality) Current work consider all the black box algorithm for the data (X_i, Y_i) with domain $\mathbb{R}^d \times \mathbb{R}$ for some constant $d \geq 1$ and shows the Binomial test achieves the optimal power. In practice, the domain $\mathbb{R}^d$ and $\mathbb{R}$ may be two wide. For example, in classification problem, the response Y is supported on a discrete set. The Theorem 2.1 still holds while the optimal power in Theorem 2.2 may be larger due to the more constrained space for black box algorithms. The optimality of the Binomial test may not hold with extra constraint on P. Similarly, if we have more information on P such as polynomial relationship between X and Y and the marginal distributions of X, Y, the optimal power in Theorem 2.2 may also increase.

2. (Symmetric) This work focuses in the symmetric algorithm. This excludes some online methods like reinforcement learning which updates the algorithm after given a new data point. It may be interesting on how to extend the stability testing to these non-symmetric algorithms.

3. (I.i.d sample) This work relies on the i.i.d. setting of the data. However, in practice, it is nearly impossible for us to have i.i.d. sample. Can we extend the method with identically distributed data without independence? Can we relax to i.i.d. setting to exchangeable data; i.e., the order of the data points does not affect the joint distribution?

4. (Inference problem) This works aims to the test the algorithm stability in the prediction. It may be interesting to check whether we can extend the test for inference problem, such as clustering, matching, and parameter estimation. The algorithm stability should be re-defined in this case since inference problem does not focus on the performance with a new data X_{n+1} but care about the stability on current (training) data $\{X_i, Y_i\}_{i\in[n]}$ with small perturbation. For example, in clustering problem, let $\hat{z}_n$ and $\hat{z}_{n-1}$ denote the estimated assignment with data $\{X_i, Y_i\}_{i\in[n]}$ and $\{X_i, Y_i\}_{i\in[n-1]}$, respectively. We define the difference

$$\Delta = \frac{1}{n-1} \min_{\pi\in\Pi_{n-1}} \sum_{i\in[n-1]} \mathbb{1}\{\hat{z}_n(i) \neq \pi \circ \hat{z}_{n-1}(i)\},$$

where Π_{n-1} is the collection of all permutation on $[n-1]$ and compare the difference Δ with ϵ.

References