Semi-supervised Active Linear Regression

Devvrit*[1], Nived Rajaraman*[2], Pranjal Awasthi[3]

[1]University of Texas at Austin
[2]University of California, Berkeley
[3]Google Research & Rutgers University

(*Equal Contribution)

NeurIPS 2022

Linear Regression

Problem: Given n points in d dimension $X \in \mathbb{R}^{n \times d}$ and corresponding labels $Y \in \mathbb{R}$, we want to find $\beta^* \in \mathbb{R}^d$ such that

$$\beta^* = \min_{\beta} \|X\beta - Y\|_2^2 \tag{1}$$

$\rightarrow$ Supervised learning assumes that Y is observed.

$\rightarrow$ Can be costly to get the labels of all points.

How to reduce the sample cost

Two approaches for reducing the sample complexity, that have received much attention in the contemporary ML literature:

Active Learning: The dataset is unlabeled, and the algorithm can adaptively query the labels of a small subset of data points to carry out the task.

Semi-supervised Learning: The learner has access to massive amounts of unlabeled data in addition to some labeled data, and algorithms leverage both to carry out the learning task.

Semi-supervised Active Linear Regression (SSAR)

In this work, we introduce *Semi-supervised Active Linear Regression (SSAR)*, which combines elements of both active learning and semi-supervised learning.

Problem (Agnostic SSAR)

The learner has n_{un} unlabeled points and n_{lab} points labeled a-priori in $\mathbb{R}^d$ collected in the matrix X. Denote the true labels by $Y \in \mathbb{R}^{n_{un}+n_{lab}}$. The objective is to find $\hat{\beta} \in \mathbb{R}^d$ such that

$$\|X\hat{\beta} - Y\|_2^2 \leq (1+\epsilon) \min_{\beta} \|X\beta - Y\|_2^2, \tag{2}$$

while querying the labels of as few unlabeled points as possible.

Special cases of SSAR

The SSAR problem generalizes two known problems from the literature

Active ridge-regression: The active ridge regression objective is $\|X_{\text{un}}\beta - Y_{\text{un}}\|_2^2 + \lambda\|\beta\|_2^2$. The unlabeled dataset has d points, $\{\sqrt{\lambda}e_i, i = 1, \cdots, d\}$ with corresponding labels 0.

Active kernel ridge regression: Similar to above, the kernel matrix can be augmented with the basis vectors, with the corresponding labels being 0.

Contribution

1. We introduce an instance-dependent parameter called the *reduced rank*, denoted by R_X.
   $\rightarrow$ For ridge regression, R_X is the "statistical dimension" sd_λ. **?**
   $\rightarrow$ For kernel ridge regression, R_X is the "effective dimension" d_λ **?**.

2. When $\epsilon \in (0, 1)$, we provide an algorithm with sample complexity of $O(\mathrm{R}_X/\epsilon)$ for SSAR.

3. Prove a matching *instance-wise* lower bound of $\Omega(\mathrm{R}_X/\epsilon)$ on the query complexity of any algorithm for a distributional/noisy version of the problem for the same range of ϵ.

Reduced Rank

Reduced Rank (R_X): A parameter that inuitively measures how informative the labeled dataset X_{lab} is compared to the overall dataset $X = \begin{bmatrix} X_{un} \\ X_{lab} \end{bmatrix}$.

$$\mathrm{R}_X = Tr\left(\left(X^T X\right)^{-1} X_{un}^T X_{un}\right) \tag{3}$$

$\rightarrow$ The reduced rank is always upper bounded by d.

ASURA (Active semi-SUpervised Regression Algorithm)

High-level description: The algorithm samples a subset of $m = \frac{\mathrm{R}_X}{\epsilon}$ points from X, and corresponding weights $\{w_1, \cdots, w_m\}$, and performs weighted least square regression. Namely,

$$\hat{\beta} = \min_{\beta} \sum_{i=1}^{m} w_i \left(x_i^T \beta - y_i \right)^2 \tag{4}$$

ϵ-well balanced procedure

The algorithm builds on the spectral sparsification based sampling primitive developed in **?**. We design a novel spectral sparsification mechanism which samples points sequentially and guarantees that the number of labeled points sampled is upper bounded by $\frac{\mathrm{R}_X}{\epsilon}$ with probability 1.

$\rightarrow$ The randomized BSS algorithm **?** gives only a probabilistic bound on the total (unlabeled + labeled) number of points sampled.

$\rightarrow$ This is not sufficient to bound the number of unlabeled points sampled - the number of points sampled can be correlated with the nature of the points sampled (i.e. labeled/unlabeled).

$\rightarrow$ Our algorithm sidesteps having to deal with these correlations.

Lower Bound

Distributional SSAR: Labels revealed to the learner are corrupted by independent noise, as $y = f(x) + Z$, where the noise $Z \sim \mathcal{N}(0, \sigma_x^2)$. The objective is to minimize the generalization error,

$$\mathbb{E}\left[\frac{1}{|X|}\sum\nolimits_{x\in X}\left(\langle\widehat{\beta}, x\rangle - f(x)\right)^2\right] \tag{5}$$

$\rightarrow$ It is a special case of agnostic SSAR.

Lower Bound (contd.)

Theorem (Lower bound)

Suppose $\epsilon \in (0,1)$. *In distributional SSAR, for each* X *and learner there exists an instance where the learner must query* $\Omega(\frac{\mathrm{R}_X}{\epsilon})$ *labels to guarantee,*

$$\mathbb{E}\left[\|X\widehat{\beta} - f(X)\|_2^2\right] \leq (1+\epsilon) \min_{\beta \in \mathbb{R}^d} \mathbb{E}\left[\|X\beta - f(X)\|_2^2\right]$$

.

$\rightarrow$ Reduced rank characterizes the sample complexity on a per-instance basis for distributional SSAR.

Conclusion

$\rightarrow$ We show that the sample complexity of distributional SSAR is characterized on a per-instance basis by a new parameter known as the reduced rank, R_X. The sample complexity is shown to be $O(\frac{\mathrm{R}_X}{\epsilon})$ for $\epsilon \in (0, 1)$.

$\rightarrow$ For ridge regression, $\mathrm{R}_X = \mathrm{sd}_\lambda$ (statistical dimension) and for kernel ridge regression, $\mathrm{R}_X = d_\lambda$ (effective dimension), resulting in a sample complexity of $\frac{\mathrm{sd}_\lambda}{\epsilon}$ for the active ridge regression, and $\frac{d_\lambda}{\epsilon}$ for the active kernel ridge regression problem.