Training Data Subset Selection for Regression With Controlled Generalization Error

200050031, 213050036, 18B030021

May 2022

1 Abstract

In this report, we aim to replicate the results presented in [1]. We shall here consider the problem of data subset selection for the L2 regularized regression problem which seek to minimize the training loss with respect to both the trainable parameters and the subset of training data, subject to error bounds on the validation set. Our aim is to first find trainable parameters by fixing the subset and then minimising the subset based on trainable parameters. In order to do that, first we will define our problem formulation using dual of the original training problem with simplified constraint and then we will show that dual of the problem is monotone and α-submodular.

2 Problem formulation

2.1 Data selection

Consider a problem of linear regression with training dataset $(x_i, y_i)_{i\in D}$ and validation dataset $(x_j, y_j)_{j\in V}$ partitioned as V=$\cup_{q\in[Q]}V_q$ with model class h_w. The minimization problem is then formulated as follows with L_2 regularized training loss, subject to constraints that bounds mean squared loss(MSQ)-

$$minimise_{S\subset D,w} \textstyle\sum_{i\in S}[\lambda\|w\|^2 + (y_i - h_w(x_i))^2],$$
$$\text{subject to } \frac{\sum_{j\in V_q}(y_j - h_w(x_j))^2}{|V_q|} \leq \delta, \forall q \in [Q], |S| = k$$

where λ is coefficient of regularizer; cardinality constraint limits the number of training samples to be chosen, and validation wrror constraint ensure that predictor's loss remain below some δ.

2.2 Soft-constraint approach

Introduce new slack variables $\xi_1, \xi_2, .., \xi_Q$ and replace each hard validation constraint with soft constraint i.e. $\frac{1}{|V_q|}(y_j - h_w(x_j))^2 \leq \delta + \xi_q$. Now the optimization problem becomes-

$$minimise_{S \subset D, w, \{\xi_g\}_{q \in [Q]}} \sum_{i \in S}[\lambda \|w\|^2 + (y_i - h_w(x_i))^2] + C \sum_{q \in V_q} \xi_q,$$
$$\text{subject to } \frac{\sum_{j \in V_q}(y_j - h_w(x_j))^2}{|V_q|} \leq \delta + \xi, \forall q \in [Q],_q \geq 0, \forall q \in [Q] and |S| = k$$

For any given set S, let the optimal value of the parameters be $w^*(S)$ and $\xi_q^*(S)$. Then we can formulate data selection problem as-

$$\text{g(S)} = \sum_{i \in S}[\lambda \|w^*(S)\|^2 + (y_i - h_{w^*(S)}(x_i))^2] + C \sum_{q \in [Q]} \xi_q^*(S)$$

2.3 Simplified constraint approach

Proposition 1: Given a fixed training set S, let $\mu = [\mu_q]_{q \in [Q]}$ be the Lagrangian multipliers for the constraints $\{\frac{1}{|V_q|} \sum_{j \in V_q (y_j - h_w(x_j))^2 \leq \delta + \xi_q}\}_{q \in [Q]}$ then,

$$F(w, \mu, S) = \sum_{i \in S}[\lambda \|w\|^2 + (y_i - h_w(x_i))^2] + \sum_{q \in [Q]} \mu_q [\frac{\sum_{j \in V_q}(y_j - h_w(x_j))^2}{|V_q|} - \delta]$$

Relation between f(S) and g(S). Given a fixed S, the optimization problems f(S) and g(S) are equivalent for convex losses. However, for non-convex function, f(S) would serve as a lower bound for g(S).

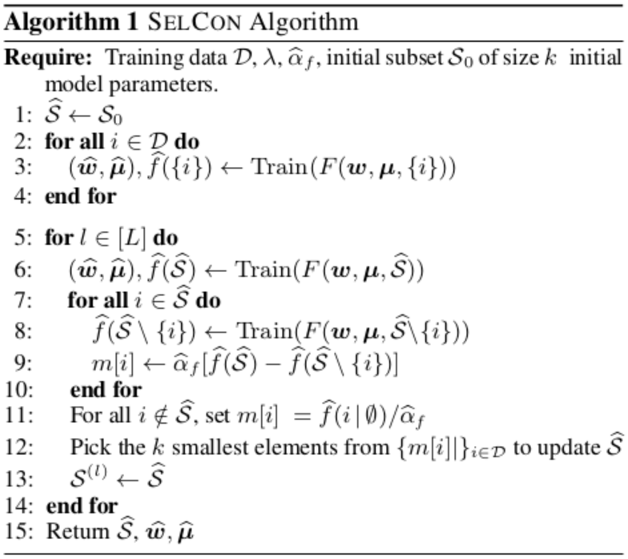
3 Characteristic of f(S)

1. f(s) is monotone meaning f(a|S) $\geq$ 0 for all S$\subset$D and a$\in D \setminus S$
2. f(S) is α-submodular meaning f(a|S)$\geq$ αf(a|T) for S$\subseteq$T and a$\in$D$\setminus$T with submodularity parameter α $\geq$0 and when α=1, f(S) will be submodular
3. Given a set S, the generalized curvature of f(S) is defined as-
$\kappa_f(S) = 1 - min_{a \in D} \frac{f(a|S \setminus \{a\})}{f(a|\phi)}$

4 Introduction to SELCON

SELCON is an iterative majorization-minimization algorithm for data subset selection, that works for α-submodular functions and enjoys an approximation guarantee even when the training provides an imperfect estimate of the trained model.

4.1 SELCON Algorithm



Algorithm 1 SELCON Algorithm

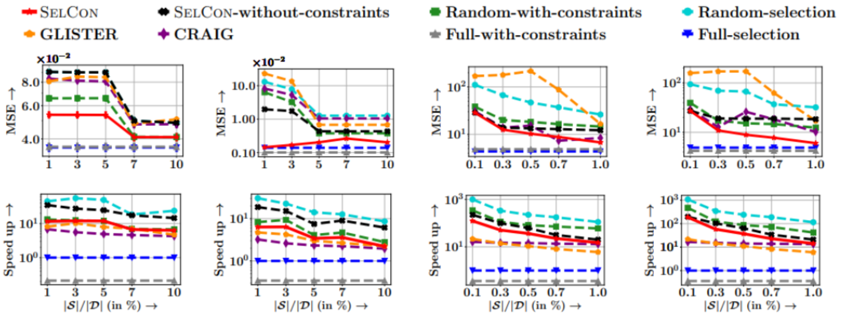
```
Require: Training data $\mathcal{D}$, $\lambda$, $\widehat{\alpha}_f$, initial subset $\mathcal{S}_0$ of size $k$ initial
    model parameters.
$\widehat{\mathcal{S}} \leftarrow \mathcal{S}_0$
for all $i \in \mathcal{D}$ do
    $(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\mu}}), \widehat{f}(\{i\}) \leftarrow \text{Train}(F(\boldsymbol{w}, \boldsymbol{\mu}, \{i\}))$
end for
for $l \in [L]$ do
    $(\widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\mu}}), \widehat{f}(\widehat{\mathcal{S}}) \leftarrow \text{Train}(F(\boldsymbol{w}, \boldsymbol{\mu}, \widehat{\mathcal{S}}))$
    for all $i \in \widehat{\mathcal{S}}$ do
        $\widehat{f}(\widehat{\mathcal{S}} \setminus \{i\}) \leftarrow \text{Train}(F(\boldsymbol{w}, \boldsymbol{\mu}, \widehat{\mathcal{S}} \setminus \{i\}))$
        $m[i] \leftarrow \widehat{\alpha}_f[\widehat{f}(\widehat{\mathcal{S}}) - \widehat{f}(\widehat{\mathcal{S}} \setminus \{i\})]$
    end for
    For all $i \notin \widehat{\mathcal{S}}$, set $m[i] = \widehat{f}(i \,|\, \emptyset)/\widehat{\alpha}_f$
    Pick the $k$ smallest elements from $\{m[i]|\}_{i\in\mathcal{D}}$ to update $\widehat{\mathcal{S}}$
    $\mathcal{S}^{(l)} \leftarrow \widehat{\mathcal{S}}$
end for
Return $\widehat{\mathcal{S}}, \widehat{\boldsymbol{w}}, \widehat{\boldsymbol{\mu}}$
```

Explanation of algorithm:

5 Experiment

The paper use two models— a simple linear regression model and a two layer neural network that consists of a linear layer of 5 hidden nodes and a ReLU activation unit. In all our experiments, we use a learning rate of 0.01. We choose the value of as the 30

6 Result



(a) Cadata (b) Law (c) NYSE-High (d) NYSE-Close

7 Conclusion

We replicated the results presented in [1] and found that SELCON was the best performer among all the other algorithms.

8 References

Durga, S., Rishabh Iyer, Ganesh Ramakrishnan, and Abir De. ”Training Data Subset Selection for Regression with Controlled Generalization Error.” In International Conference on Machine Learning, pp. 9202-9212. PMLR, 2021.