Codes Reproducibility README for “Multi-resolution subsampling for linear classification with massive dataset”

This file is designed to help readers reproduce the main result in the paper.

Codes Directory

Below is a list of main files contained in the *Codes\_Reproducibility.zip* file:

1. **basic** This file folder includes some basic and useful functions.

DataGenerator.R: To generate the synthetic data for Figure 1, Figure 2, and the simulation study.

getIBOSS.R: To implement IBOSS in Cheng et al. (2020).

getMLE.R: To get the MLE estimation of a weighted logistic regression.

getsSVM.R: To solve a squared SVM.

PA.R: To calculate prediction accuracy.

Solvers.R: This file contains all implementation functions of our proposed method and some existing subsampling methods(OSMAC, IBOSS, MSCLE).

1. **Rcpp** This file folder contains the implementations of our proposed method, uniform sampling, and OSMAC by C++, which is used to analyze the computing time.

OSMAC.cpp: OSMAC for logistic regression.

OSMAC\_dwd.cpp: OSMAC for DWD.

RBIMP.cpp: the proposed method for logistic regression.

RBIMP\_dwd.cpp: the proposed method for DWD.

UNIF.cpp: uniform sampling.

1. **realdata** This file folder contains some codes for preprocessing real datasets. Each “XXX\_pre.R” file generates a “XXX\_pre.Rdata” file, which contains the preprocessed data of dataset named “XXX”.

covertype\_pre.R: preprocessing for forest cover type dataset.

PM25\_pre.R: preprocessing for Beijing multi-site air quality dataset.

SUSY\_pre.R: preprocessing for supersymmetric benchmark dataset.

1. **Figure1.R** This file generates Figure 1 in the paper.
2. **Figure2.R** This file generates Figure 2 in the paper.
3. **Figure3&4.R** This file conducts the simulation comparison of UNIF, OSMAC, IBOSS, MSCLE, and our proposed method, for logistic regression, in terms of MSE, with different subsample sizes *r* and fixed full sample size *n*. The results are shown in our paper Figure 3 and Figure 4. This file provides an example of the code to recreate only the results of Case 1 in Figure 3 and Figure 4. However, one can easily reproduce all the results in Section 5.1 by setting the value of “Case” in line 63 of this file to be within {1,2,3,4}.
4. **FigureS1.R** This file includes the codes of sensitivity analysis of Cr0 for our proposed method, for logistic regression, with different subsample sizes *r* and fixed full sample size *n*. The results are shown in Figure S.1 in the supplementary material of our paper. Similar to *Figure3&4.R*, this file provides an example of the code to recreate only the results of Case 1 in Figure S.1. One can reproduce all the results in Section S.12.1 by setting the value of “Case” in line 63 of this file to be within {1,2,3,4}.
5. **FigureS2&S3.R** This file includes the codes of simulation comparison of UNIF, OSMAC, and our proposed method, for DWD, in terms of MSE, with different subsample sizes *r* and fixed full sample size *n*. The results are shown in Figure S.2 and Figure S.3 in the supplementary material of our paper. Similar to *Figure3&4.R*, this file provides an example of the code to recreate only the results of Case 1 in Figure S.2 and Figure S.3. One can reproduce all the results in Section S.12.3 by setting the value of “Case” in the line 62 of this file to be within {1,2,3,4}.
6. **Table2\_95CI.R** This file takes Case 1 for logistic regression as an example and calculates the empirical coverage probability and the average length of the confidence interval for the intercept and the first component of the slope of the predictor The results are shown in Table S.2.
7. **Computing\_time** This file folder contains codes of the simulation comparison of computing time between our proposed method and OSMAC via Rcpp. The result is presented in Table 1 and Table S.3. (Different computing environments may result in different computing times. The computations for results in Table 1 and Table S.3 were both carried out on a laptop running Windows 10 with an Intel I5 processor and 8GB memory.)

Time\_Full.R: generating the results of the last row “Full” in Table 1 and Table S.3.

Time\_logistic.R: generating the results of the column “d = 21” in Table 1.

Time\_dwd.R: generating the results of the column “d = 21” in Table S.3.

1. **Sampling\_time** This file folder contains codes of the simulation comparison of computing time for the sampling step between our proposed method and OSMAC via Rcpp. The result is presented in Table S.4.

TableS4\_sim\_log.R: generating the results of the row “logistic regression” for “d = 21” in Table S.4.

TableS4\_sim\_dwd.R: generating the results of the row “DWD” for “d = 21” in Table S.4.

TableS4\_SUSY\_log.R: generating the results of the row “logistic regression” for “d = 18” in Table S.4.

TableS4\_SUSY\_dwd.R: generating the results of the row “DWD” for “d = 18” in Table S.4.

1. **realdata\_logistic.R** This file provides codes to compare UNIF, OSMAC, MSCLE, and our proposed method, for logistic regression, in terms of MSE and prediction accuracy, with different subsample sizes *r*. The results for different real data sets are shown in Figure 5, Figure S.5 (top panel) and Figure S.6 (top panel).
2. **realdata\_dwd.R** This file provides codes to compare UNIF, OSMAC, and our proposed method, for DWD, in terms of MSE and prediction accuracy, with different subsample sizes *r.* The results for different real data sets are shown in Figure S.4, Figure S.5 (bottom panel), and Figure S.6 (bottom panel).
3. **Reply** This file folder contains codes for reproducing part of the results in the replies to the referees.

RE1Q6.R: generating the results of Table R1 in the reply to Referee 1’s comment 6.

RE2Q9.R: generating the results of Figure R2 in the reply to Referee 2’s comment 9.

GDgetMLE.R, IRLSgetMLE.R: to implement gradient decent method and IRLS for logistic regression in *RE1Q6.R*.

To use *realdata\_logistic.R*(or *realdata\_dwd.R*, *TableS4\_SUSY\_log.R*, *TableS4\_SUSY\_DWD.R*), one should first generate the preprocessed Rdata file with codes in the folder *realdata* (for example, for the supersymmetric benchmark dataset, the code *SUSY\_pre.R* generates the preprocessed Rdata file *SUSY\_pre.Rdata*), then load it in the *realdata\_logistic.R*(or *realdata\_dwd.R*, *TableS4\_SUSY\_log.R*, *TableS4\_SUSY\_DWD.R*).

Variables

Some variables are named differently in the codes from our paper. For readers’ convenience, we list some important variables in the codes.

1. **N** The full sample size, i.e. *n* in the paper.
2. **p** The dimension of covariates (without intercept term), i.e. *d* minus 1 in the paper.
3. **n0** The size of , i.e. *r0*in the paper.
4. **ns** A sequence of different sampling sizes *r*.
5. **Cs** A sequence of *.*