Random Bits Regression

Author: Yi Wang
27/Feb/2014

Introduction:
The software requires modern CPUs with **SSE** instructions. It contains two programs: **rbr_model** is used to build a model on training dataset and **rbr_predict** is stream based real time predictor on testing dataset.

Usage:

rbr_model *[options]* ***trainX trainY model***

Input Format:

trainX should be comma delimited or colon delimited or tab delimited text files. The data is a dense matrix without header. Each line represents a sample and each column represents a variable. Missing value is define as |X|>=1e15. Missing value will be replace with mean value automatically.

trainY should be the same format as trainX. If a trainY variable contains only 0,1 values, it will be automatically treated as binary variables and logistic regression will be applied instead of least square regression.

Output Format:

The output model file is a human readable text file. The first line is variable importance weights. The second part is the mean and precision of each variable. The third part is regression setting. The last part is the random bits' weight, threshold and twisted features.

Options:

-b

Number of random bits used. The default setting (100,000 bits) balances well on various datasets and the parameter is usually not tuned. For very large datasets, b=10,000 can be used.

-r

Scaled L2 regularization parameter. The default setting is 0.01. A bit tuning of this parameter is encouraged, sometime leads to better predictions. Larger value leads to under-fitting and small value leads to over-fitting.

-t

Number of features twist in one bit. The default setting is 2 features. A bit tuning of this parameter is encouraged, sometimes leads to better predictions.

-c

Number of L-BFGS corrections. The default value (256 corrections) balances well on various datasets and the parameter is usually not tuned. Small value leads to less memory usage and longer convergence time.

-m

Maximum of L-BFGS iteration. The default value is zero which means no limits.

-T

Number of threads. The default value is 0 which means using all CPU cores.

-w

A file contains the importance of each variable. This is used in variable selection/weighting. 0 means ignore the corresponding variable. The model file generated by last rbr_model run can be feed into the parameters. This iteration can be repeated and leads to a feature selection method.

Usage:

rbr_predict ***model <testX >yhat***

rbr_predict is a stream based real time predictor. This command will predict testX according to model and save the prediction to yhat.

Citation:

Yi Wang[*], Yi Li[*], Momiao Xiong, Li Jin[#], Random Bits Regression, A Strong General Predictor.

Enjoy!