Supplementary Information:
Stability Selection Algorithm

July 2021

Stability Selection is an algorithm aiming to enhance and improve the existing feature selection algorithm through combining it with sub-sampling. The basic idea behind it is:

- First, introduce more noises to the original problem by drawing bootstrap sample of the data;
- Then, use the given feature selection algorithm to determine which features are important in every sampled version of the data;
- Next, aggregate the results on each bootstrap sample to get a *stability score* for each feature in the data;
- Finally, select features by choosing an appropriate threshold for the stability scores.

It can works in high-dimensional data setting, and it can provide control for some error rates of false discoveries in the finite sample setting.

Given a dataset denoted by $\mathbf{X}^{n\times p}$. Suppose that in the learning algorithm (or feature selection algorithm), there is a parameter λ that controls the strength of regularization in the problem; in addition, for each value of λ, we get a set $\hat{S}^{\lambda}$, which indicates which variables to select. Then, the Stability Selection algorithm works as followed:

(1) Define a candidate set of regularization parameters, denoted by Λ and a subsample number N

(2) for each value λ in the set Λ:

  (a) Start with the full dataset $\mathbf{X}$

  (b) For each i in $1, ..., N$:

    (i) Subsample from $\mathbf{X}^{n\times p}$ without replacement to generate a smaller dataset of size $\lfloor n/2 \rfloor$, denoted by $\mathbf{X}_{(i)}$

    (ii) Run the selection algorithm on the sample dataset $\mathbf{X}_{(i)}$ with regularization parameter λ to obtain a selection set $\hat{S}^{\lambda}_{(i)}$

  (c) Given the selection sets from each subsample, calculate the empirical selection probability for each model component, i.e. for $k = 1, ...p$:

$$\hat{\Pi}^{\lambda}_{k} = \mathbb{P}[k \in \hat{S}^{\lambda}_{(i)}] = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}_{k \in \hat{S}^{\lambda}_{(i)}}$$

  The selection probability for component k is its probability of being selected by the algorithm.

(3) Now, given the selection probabilities for each component and for each value of λ, construct the stable set according to the following definition:

$$\hat{S}^{\text{stable}}_{(i)} = \{k : \max_{\lambda\in\Lambda} \hat{\Pi}^{\lambda}_{k} \geq \pi_{\text{thr}}\}$$

where π_{thr} is a predefined threshold.

Note that the algorithm does not simply select the best $\lambda \in \Lambda$ and then use it in the algorithm. Instead, it finds a set of "stable" variables that are selected with high probability.

References

[1] Han Liu, John Mulvey, and Tianqi Zhao. A semiparametric graphical modelling approach for large-scale equity selection. *Quantitative finance*, 16(7):1053–1067, 2016.

[2] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

[3] Shiqiong Huang and Micol Marchetti-Bowick. Summary and discussion of: Stability Selection. `https://www.stat.cmu.edu/~ryantibs/journalclub/stability.pdf`. Online, Accessed: 2021–07-01.

[4] Welcome to stability-selection's documentation. `https://thuijskens.github.io/stability-selection/docs/index.html`. Accessed: 2021–07-01.