Sensitivity Analysis on oat data (ELY keskus)

Paavo Reinikka

27 August 2019

Setup for SA

Sensitivity analysis was performed on full oat data, and the focus of the analysis was on the input-output sensitivity of various elnet models (as defined and implemented in the r package **glmnet**). The models were differentiated on two axis: *alpha* being the tuning parameter for the continuum Ridge - Lasso, and *lambda* being the parameter deciding the amount of penalization (regularization). The parameters were defined on a discrete grid: $\lambda \cup \alpha$ st. $\lambda \in [min, 1se]$ and $\alpha \in [0, 0.25, 0.5, 0.75, 1]$, where *lambda* value min corresponds to model with minimum cv error (per *alpha*) and value 1se corresponds to maximum cv error *within* 1 se from that minimum. Different model runs are therefore consistent not by the exact values of lambda, but rather the heuristic for choosing the lambda. This affords us some added aggregation over the data.

The SA performed on the oat data varied only predictors (input-output SA), and only after model had been built; i.e., effects of variations/perturbations in the training data on the modelling (building of the model) were not observed. Also, modelling assumptions were not varied directly, as a source of output variation. However, some insights to the effects of the main elnet parameters were acquired indirectly by performing input-output SA along different values of alpha, lambda.

Used measures and their limitations

In the SA literature there are many different sensitivity measures, both local and global, with varying degrees of constraints. Use of *sigma-normalized coefficients* (Saltelli et.al.[1]) are considered a sufficient measure when model is known to be linear and predictors un-correlated. In a more general setting however, when a certain level of agnostisism is preferred with regards to the model and the input space (e.g., ML setting), MC sampling coupled with either variance or enthropy based measures are preferred. SA here is performed using r package *sensitivity* and as sufficient measures Sobol's 1st and total order indices[2].

For model $y = f(\mathsf{x}_1, ..., \mathsf{x}_p)$ where y is a scalar ('production') and x's are the columns of the input matrix, the 1st order sensitivity index, corresponding to *main effects* of x_i, is defined:

$$\mathsf{S}_i = \frac{\mathsf{V}_{\mathsf{x}_i}(\mathsf{E}_{\mathsf{X}_{-i}}(\mathsf{y}|\mathsf{x}_i))}{\mathsf{V}(\mathsf{y})}$$

where the inner (x_i conditional) expectation is taken over all other predictors, and the outer variance over x_i. Similarly the total order index, which corresponds to all other effects of x_i (through it's interactions with other variables) can be defined:

$$\mathsf{ST}_i = \frac{\mathsf{E}_{\mathsf{X}_{-i}}(\mathsf{V}_{\mathsf{x}_i}(\mathsf{y}|\mathsf{X}_{-i}))}{\mathsf{V}(\mathsf{y})} = 1 - \frac{\mathsf{V}_{\mathsf{X}_{-i}}(\mathsf{E}_{\mathsf{x}_i}(\mathsf{y}|\mathsf{X}_{-i}))}{\mathsf{V}(\mathsf{y})}$$

These indices are often presented as for independent input distributions, when the property $0 \leq \mathsf{S}_i \leq \mathsf{ST}_i \leq 1$ holds.

Relevant literature

[1]Main source by Saltelli et. al. http://www.andreasaltelli.eu/file/repository/A_Saltelli_Marco_Ratto_Terry_Andres_Francesca_Campolongo_Jessica_Cariboni_Debora_Gatelli_Michaela_Saisana_Stefano_Tarantola_Global_Sensitivity_Analysis_The_Primer_Wiley_Interscience_2008_.pdf

[2]I M Sobol'. Sensitivity estimates for nonlinear mathematical models.Mathematical Modelling andComputational Experiments, 1:407–414, 1993.