Is Random Forest the only algorithm to measure the importance of input variables …?

I have three time series say (Stock price open, Stock price high, Stock price low) and one output (Stock price close) and I need to know which of the 3 inputs has a greater effect on my output. R's Random Forests' (importance) ' %IncMSE yields the importance for this case.

Are there any other algorithms apart from Random Forest, to measure the importance of a input variable?

r machine-learning random-forest predictor

edited Feb 17 at 14:40

General Abrial
15.6k 4 35 74

asked Feb 17 at 11:31

Sudharsan
27 5

Gradient Boosting Machines with the "gbm" R package is other algorithm with similar performance than RF, and showing also the importance of the predictors, but the ranking should be very similar – Jesus Herranz Valera Feb 17 at 11:44

Maybe `Boruta algorithm` will help you: cran.r-project.org/web/packages/Boruta/Boruta.pdf ? – Maju116 Feb 17 at 11:50

1 Ordinary regression coefficients will tell you "importance" in a specific sense. What kind of importance are you interested in? – General Abrial Feb 17 at 14:39

2 @user777 Actually and unless the predictors have been standardized (never a good practice), your statement is not true since regression coefficients are expressed in the unit of the predictor. As such they are *not* scale invariant and sensitive to variations in the moments. A better, quick and dirty heuristic is to rank the standardized metrics such as F-statistics, t-values and/or chi-squares associated with the parameters, as appropriate. For the "state-of-the-art" in deriving relative importance, see Ulrike Groemping's papers on 'relaimpo'...a multivariate approach to relative rankings – DJohnson Feb 17 at 14:52

@Maju116 I just now saw Random uniform forest package in R . I think it does the best with this use case of finding the input variable affecting the response variable . Also if the input is categorical, say region(A,B,C) it also shows the importance of each region . – Sudharsan Apr 26 at 11:32

1 Answer

Any supervised regression/classification model, that I can think of, could be bootstrap aggregated (bagged) and therefore variable importance could be computed. It would just be a little slow to train e.g. svm 50 times compared to growing 50 trees.

> *"I need to know which of the 3 inputs has a greater effect on my output"*

I would abstain from causal interpretation of importance and at most see it as a source of inspiration. Importance only describe usefulness of features to predict, given all other features and one specific model. Two highly redundant features will roughly share the same fixed amount of variable importance. Two features can be complimentary and have a higher variable importance, than if one of them were never included in the training. This happens if an interaction between two features is useful to predict the output.

So importance is not an universal metric, and the answer will depend on your model and and the included features. You may want to ask instead:

"Which overall relationship between input and output has e.g. an RF/SVM/NN model captured?" -You could use some fancy plots exploring the high dimensional model structure. But I can reveal that the effective RF, SVM or NN model will be something very boring like $close_t = (open_t + high_t + low_t)/3$

"Is this relationship trivial or inspiring?" - In this case, quite trivial, as the future absolute price is highly dependent on the past price. If some asset were 10€ yesterday, its probably gonna be priced 9€ or 12€ today, not 1 cent or 5000€.

Try use rolling window to predict the change of price, that is in contrary more challenging. If you were to succeed better than others, then the effective structure of a well performing empirical model could be very inspiring to form new hypotheses.

answered Feb 19 at 16:41

Soren Havelund Welling
2,871 5 17

Add Another Answer