Manish Tripathi <mtripathi@pivotal.io>

Thursday Thunder :D

Manish Tripathi <mtripathi@pivotal.io> Fri, Oct 2, 2015 at 10:53 AM
To: Srivatsan Ramanujam <sramanujam@pivotal.io>, Sarah Aerni <saerni@pivotal.io>, Gautam Muralidhar <gmuralidhar@pivotal.io>, Rashmi Raghu <rraghu@pivotal.io>, Swati Soni <ssoni@pivotal.io>, Woo Jae Jung <wjung@pivotal.io>, Esther Allas <evasiete@pivotal.io>, Anirudh Kondaveeti <akondaveeti@pivotal.io>, Ambarish Joshi <ajoshi@pivotal.io>, April Song <asong@pivotal.io>, Noelle Sio <nsio@pivotal.io>, PDS Amer <pds-amer@pivotal.io>

So after a long hiatus and just work, I thought it's high time I have troubled people again.

So here is what I have decided now. Every Thursday I would shamelessly send one email in this Thunder series :) with some question. Will wait for one week for anyone to respond else I would try to figure out the answer somehow and share it. Hopefully this shameless act of mine would make some guys to start participating for more learning.

Answer for last question:

> 1). Does low or negligible correlation among the features is a good measure of not having multi-collinearity issue?. Like many times people remove or drop a few variables and get rid of correlation. Or one of the first measure people do is get the covariance matrix of features and check which variables are correlated and then take a call on multi-collinearity. Is this a good indicator?
>
> It may sound that correlation is the reason for multi-collinearity , it is not the whole truth. Correlation is not what causes multi-collinearity. If there is correlation there will be multi-collinearity but if there is multi-collinearity that doesn't mean there would always be correlation. Multi-collinearity is a result of having a design matrix which has at least one feature a linear combination of other feature. In other words, a feature which is dependent on other features. But then isn't a dependent variable is going to have correlation also?. Not always.
>
> Consider a random variable X and another random variable Y such that
>
> Y=X^2.
>
> Y is dependent on X so in that sense there is dependence but the correlation is zero. Why?. Because correlation is NOT a measure of anything but LINEAR dependence between variables. So when you create a correlation matrix of features to check which variables are correlated, you might end up seeing zero in some cases and still there would be multi-collinearity. It doesnt happen often but theoretically it's possible.
>
> 2). ML junta doesn't seem like care about multi-collinearity. Any thoughts why?. I think I know why but good to know other people thoughts too.
>
> Because Machine learning is more about prediction rather than getting inference or relationship between variables. Having multi-collinearity is not going to affect the prediction interval. It affects the variances of the weights which get inflated. So one would have trouble in explaining the variance in outcome because of individual independent variables but it would not have any affect on prediction interval.

> And btw, Ridge and Lasso regression does alleviate the problem of multi-collinearity , so there some of this issue getting tackled by ML guys too. :D
>
> Next question would come soon. Be ready :D

[Quoted text hidden]