Statistical inference for exploratory data analysis and model diagnostics

Aarathy Babu

Exploratory data analysis is roughly associated with what we do to raw data before fitting a complex model and model diagnostics is what we do to the transofrmed data after a model is fitted.

Discoveries as rejections of null hypotheses

In EDA, the same null hypothesis can be rejected in favour of many possible alternatives.Similarly in model diagnostics, the fitted model can be rejected for many reasons (in standard linear models reasons could be non linearities, skew residual distributions etc).

- Pre-specification of intended discoveries :

In EDA and MD, the range of visual discoveries is not pre specified explicitly. The absence of prior specification is commonly interpreted as invalidating any inferences as post hoc fallacies.

The paper proposes that there is no need for pre-specification of discoverable features. This is visible through plots. Not only the occurence of discoveries ca be registered but also a description of their nature can also be noted.

- Control of Type 1 error (Calibration of the discovery process) :

There is a need to calibrate visual detection without resorting to the pseudo-calibration of post hoc quantitative tests tailored to the discovery.

The paper argues in favour of a protocol which is applied to null datasets drawn from null hypothesis in addition to the real dataset. Through this method, the performance is analyzed when there's nothing to discover(associated to the null hypothesis). Through this process, we calibrate the family-wise Type I error rate for the whole family of discoverable features. If we find any structure in the nullplots ,we can (i) tally the occurrences of spurious discoveries/rejections, and more specifically we can (ii) learn the most frequent types of features.

Reference distributions, null datasets and null plots

In case of visual inference, the test statistic is a plot of the data and the 'null distribution of plots' is similar to the null distribution of test statistics in conventional testing. The null datasets are sampled from the null hypothesis

$$H_0$$

.

How to reduce composite null hypothesis to sinngle reference distributions?

- conditional sampling given a statistic that is minimal sufficient under the null hypothesis.

- parametric bootstrap sampling
- Bayesian Posterior predictive sampling

When H0 consists of a more complex model where reduction with a minimal sufficient statistic is unavailable, parametric bootstrap sampling or posterior predictive sampling will generally be available.

Protocol 1 : The Rorschach

- Measures a data analyst's tendency to overinterpret plots in which there is no or only spurious structure. Provides a prior-to-analysis visual calibration training for the data analyst.
- To learn about factors that affect their tendency to see structure when in fact there is none.
- Estimates the effective family-wise Type I error rate but does not control it at a desired level.
- Exposes the data analyst to a number of null plots and tabulate the proportion of discoveries that are constructed falsely.
- The plot of the real data is inserted with known probability in a random location.

Protocol 2 : The lineup

- provides an inferentially valid test of whether what we see in a plot is really there
- Asks the observer to the identify the real data plot from among a set of decoys (null plots).
- inferentially valid p-value is the result of a lineup protocol
- A line up consists of 20 plots, where 19 of them are from null datasets and one is the real data plot which is placed at a random location.
- If the observer chooses the real data plot, the pvalue can be assigned a value of 0.05 (1/20). That is there is a 1 in 20 chance that the plot of the real data will be singled out.

Characteristics of a line up protocol :

- The viewer may simply be asked to find a plot that stands out. He or she need not poit out why it is different.
- Can be self administered by the data analyst provided that the location of the plot is unknown to the analyst but a second set of self administration will not be inferentially valid.
- The viewers could also be asked to select one of more special plots or rank them completely or partially.
- The protocol can be repeated with multiple independently recruited viewers who havent seen the plots before.

If K investigators are employed and k (k less than or equal to K) selected the plot of the real data, the combined p-value is obtained as the tail probability P(X less than or equal to k) of a binomial distribution B(K, p = 1/20). It can hence be as small as 0.05^K if all investigators picked the plot of the real data (k = K).

Points to be considered

- The wording of viewer instructions can affect the plot choices.
- Information related to the viewer's response is useful. Especially the time taken to arrive at a choice. If its quicker then the pattern is strong is not then its either that the pattern isnt strong or that the user isnt confident in their choice. Reliabilty can be caliberated by taking time into consideration.
- There are common traits that we should be aware of and expect to see from all viewers. Rorschach protocol can help in order to fix a baseline for individual.