Jonatan Cepeda Mendoza

Software Presentation

1. Choose a package, application, software that is related to data science or machine learning

Flexplot graphically Based package in R

1. Explain what it is, what it does, where it can be used

An R package that aims to address usability of current analysis software to leverage human visual processing by providing a formula-based suite of tools that simplifies and automates much of the graphical decision-making with statistical procedures in mind.

NOTES

•Will require updating to newest version of R and download Rtools

devtools::install\_github("dustinfife/flexplot")

library(flexplot)

?flexplot <- to obtain the documentation

•General flexplot formula

Y ~ a + b | c + d

•These variables take on the form of

yaxis~xaxis + color | column\_panels + row\_panels

graphics vs tables and reported statistics, visuals are important because they enhance •transparency since they provide a medium to display the data in their entirety. The audience can see, at a glance, the appropriateness of the model, the size of the effect, and the degree of uncertainty.

•Visuals improve communication between the scientific community and the public, lay people can understand relatively sophisticated statistical procedures when that information is presented graphically.

•Highlights problems with models that are masked by traditional statistical procedures, such as nonlinearity, outliers, and heteroscedasticity that can serve as a diagnostic check

•Most importantly graphics improve encoding since nearly half of the brain is devoted to visual processing, and human visual processing can encode information in as little as a tenth of a second.

•Original R code methods is difficult to produce standard errors plots that display raw data, while also scaling axes to the means and not the range of actual data resulting in visually inflated size of the effect.

•Also difficult to produce some type of graphics (e.g Skew Location plots) without a required effort.

•This package is specifically designed to remove obstacles to sound visualization, the graphics produced using flexplot were developed using empirically-derived heuristics that maximize perceptual understanding while minimizing perceptual biases

•This allows analysts to quickly shift between statistical modeling and graphical interpretation.

Flexplot is based on the following principles

•Minimize obstacles to producing graphics, with the goal to make producing graphics simple and allow the analyst to input more resources into interpretation

•Design graphics that leverage human strengths and mitigate human biases, by highlighting uncertainty, reveal whether chosen models are appropriate, and improve encoding of statistical information.

Ggplot2 vs flexplot

• GGplot2 is based on stacked layers to produce a final visualization, this quickly can add a considerable amount of code for one visualization having to apply a x/y axis layer, point layer, summary point layer and errorbar layer. Less programming experienced analysts might not abandon point-and-click software to produce these graphics due to difficulty which results in the flexibility of ggplot to not be used.

•Flexplot is simple and easy to use but is more limited than ggplot2, the goal is rather to visualize statistical models with ease and covers most graphics analysts will use for modeling. One note is that flexplot is using ggplot2 objects for the visualizations which means that it can be edited and/or layered for further customization.

•By adopting generalized linear model (GLM) grammar, flexplot adopts the convention of the general formula of y~x\_1+x\_2 which allows for a notational consistency between modeling and visualization.

•With flexplot, interacting variables and polynomials terms, the raw data are displayed exactly as they are and if interactions are present the visual will show it.

Producing univariate, bivariate and multivariate graphics in Flexplot

•Flexplot behind the scenes will visualize numeric variables as histograms and categorical as barcharts in univariate graphing. Histograms are produced using standard R conventions, barcharts on the other hand flexplot will sort the categories by sample size instead of by alphabetical order (done in ggplot, which violated the principle of relevance)

•If the categorical variable is represented as a number, applying the factor function will produce the proper barchart.

•Flexplot will automatically produce an appropriate bivariate graphic, depending on the type of predictors and outcome variable. For example, a numeric predictor and outcome will produce a scatterplot, a numeric predictor and categorical outcome a logistic curve graph, a categorical predictor and numeric outcome will result in a beeswarm plot, and categorical predictor and outcome will produce an association plot.

•Visualizing raw data can become tricky, particularly when categorical variables are involved but flexplot can handle overlapping datapoints from a categorical predictors.

•Flexplot instead jitters data proportional to the density of the data, which means regions of little density will have little jittering and high density will exhibit more jittering. This is opposed to standard practice since those methods will adjust a value slightly so points can scatter (ex, 1 will become 1.02 or .97) resulting in a beeswarm plot that will show density as a violin plot does but allows the user to differentiate between a 15k vs 15 observations due to the presence of points instead of just density.

•Flexplot one can control for the amount of jittering in multiple ways, as a Boolean, as a number or as a vector just as it is in geom\_jitter()

•Flexplot can also specify the summary statistics of the “whiskers” the default is interquartile range but can be specified to sterr (standard error) or stdev (standard deviation).

•Flexplot scatter has different options for line of fit, the default being loess

•Flexplot can model categorical outcomes such as a logistic regression (binary outcome) and utilize the ogive curves used to model the data with a simple method argument

•Flexplot can also discern between two categorical predictor and outcome vars and will generate an association plot which plots the deviation of each cell from its expected frequencies.

•Association vs Barplot, it best maps into what sorts of questions viewers are interested in asking

\*modeling association between categorical variables is traditionally done using a chi-squared test which compared observed vs expected frequencies. The association plot displays observed (height of bar) vs expected (y axis at zero) frequencies following principle of compatibility

•Related t-test models the difference between scores (ex from Time1 to Time2) Flexplot has an additional option (related = TRUE) that tells flexplot to plot the differences rather than the groups.

•Use visualize() (which uses mixed models)

•Showing raw data allows readers to determine whether the chosen model is appropriate, and it communicates the degree of uncertainty about the model. However, when there are a large number of datapoints, it increases cognitive load and masks salient characteristics. Flexplot offers three options to the overlap problem, first is to suppress raw data (not recommended but can be done). A second option is to reduce transparency which draws more attention to the fit of the model rather than raw data (which may or may not be a good thing). Third and best option is to sample, sampling allows the visual-processing system to not be overly influenced by the fit. It is important that the visual display of fit (median+IQR, loess, regression) not be estimated from the sampled data. Rather the fit should correspond to the entire dataset, flexplot does this in the background.

•Flexplot utilizes four different strategies to visualize multivariate relationships, plotting a dimension as a different color/line/shape, plotting a dimension in a row or column panels, visualizing conditional relationships with added variable plots and overlaying ghost lines

•Added Variable Plots are underused yet extremely useful, they show the relationship between a predictor of interest and the residuals of an existing model. An example is therapy type vs weight loss after controlling for motivation, the analyst could build a model predicting weight loss from motivation then residualize that relationship and then show a beeswarm plot of residuals for each type of therapy. This is what AVPs do which reduce cognitive load.

•Flexplot in addition adds the mean back into the residuals so that the y-axis retains the original scale

•Colors/Lines/Shapes/Paneling

•Binning done easier/more intuitive

•Ghost lines repeat the relationship from one panel to the other panels to make it easier to compare

•Flexplot could still produce difficult to interpret graphics when all four flexplot slots are occupied. To remedy this, apply some strategies such as specifying two bins instead of the default three, could remove confidence intervals, reduce opacity of the data, plot regression lines as well as reducing the levels of variables in the second slot or introduce ghost lines.

•Possible to use a AVP but another option is to combine visuals with statistical modeling

•Flexplot aims to automate much of the visualizations that accompany data analysis which are based on scientifically-derived principles of visual perception and promise to promote sound data analysis and enhance data decision-making.

Additional methods within flexplot, visualize(), designed to generate a graphic that matches the formula used in a fitted model (similar to summary(), coef()) it will also generate diagnostic plots. With the option to present all or individually

•With mixed models, visualize function will randomly sample from the random effect.

•Compare.fits() which uses model comparison metrics such as AIC, BIC, Bayes Factor, R^2, p-values to see how the two models differ in terms of fit. It will overlay the fit of both models onto the raw data and can take many of the same arguments such as ghost line

•Estimates() function was designed to report parameter estimates and effect sizes, which takes a fitted object as input. The method will then determine which estimates are most appropriate. Example, grouping variables will report means, mean differences, and cohen’s d, as well as 95th percentile confidence interval. While numeric variables using estimate will report the intercept, slopes and standardized slopes as well as corresponding confidence intervals. Will report R squares and semi partial R squared except when there are interactions in the model.

•Model.comparison(), similar to anova() function in base R but it includes additional estimates including AIC, BIC, and Bayes Factor (BIC-derived). It will report the quantiles of the differences in prediction. The last reported numbers indicate the maximum difference, between both models.