MYDATAMODELS – ZGP ENGINE

TRANSFORMATIVE DATA PRE-PROCESSING

Table of Contents

Introduction and Forward

This document will describe a number of factors which should be addressed when considering a data pre-processing strategy as part of an analytic pipeline. The primary emphasis will be on the factors of relevance to the application of machine learning algorithms.

For the purposes of this document, the practice of analytic data pre-processing describes the alteration of original raw data values in a dataset. This includes noise reduction, smoothing of series values, normalization, the removal of outlier data points, and transformation of scale (logarithmic, etc.).

From this definition, this document will exclude the replacement of missing values as there is a distinct difference in function and effect between this activity and the other transformative activities.

The considerations involved in the replacement of missing values are a topic worthy of their own discussion and will be discussed only superficially in this document.

The Purpose of Data Pre-Processing

There are two primary drivers for the practice of pre-processing raw values in data:

- Human interpretation
- Algorithmic imposed constraints/limitations

Human Interpretation

Transforming values, scales, and units is a completely valid process for the purpose of human comprehension and interpretation of data. However, the human interpretation of data is irrelevant to a machine learning algorithm. Each machine learning algorithm is different in the way it examines, evaluates and applies logic to produce a mathematical model describing data.

The transformations that facilitate human comprehensions are seldom the same transformations that facilitate analysis by a given arbitrary machine learning algorithm. For this reason, the pre-processing of data for human consumption should be an independent and separate effort from the pre-processing of data for machine consumption.

For the remainder of this document, the human consumption of data will be disregarded as not being relevant to pre-processing of data for machine analytic consumption.

Algorithmic Imposed Constraints/Limitations

There exist many machine learning and statistical algorithms that impose constraints upon the range of values both within a data series and across data series. It is important to note that if an algorithm requires such pre-processing in order to create an optimal model, it is a limitation of the algorithm, not a universal requirement of all analysis and modeling.

These algorithmic constraints are often immutable and will dictate the necessary transformative processes required. As such there is no general mitigation strategy to employ that alleviates these constraints.

It is important to note however, that such constraints are specific to each algorithm. The application of transformative pre-processing should only be applied in response to specific requirements by a given algorithm for optimal results. Additionally, just because one algorithm imposes certain constraints does not mean that transformations required by that algorithm are required or even beneficial when applied to data to be considered by another algorithm. For optimal results, the requirements and nature of each algorithm must be considered independently and the data prepared accordingly.

Considerations when Employing Transformative Pre-Processing

Any time raw data is transformed for the purpose of analysis, a careful consideration must be employed with regard to the impact and implications such transformations produce.

Some general considerations follow.

Imparting attributes of sample composition upon individual records

Care must be taken and consideration must be made when performing transformative pre-processing which may impart some attributes describing the full sample into each individual record. This includes actions such as normalizing with regard to sample mean or variable minima/maxima.

Such actions have numerous implications of interest.

The first is that data set partitioning of that data set into hold-outs then violates strict statistical isolation between the various sets and hold-outs. This most often leads to some degree of overly optimistic performance estimation or some degree of curve fitting specific to the original data set.

Curve fitting in this manner is particularly problematic as there is no way to detect it prior to evaluating the model against a second completely independent data set (often requiring a completely distinct collection effort to fully assess). In many cases this does not occur until model deployment. This occurs because each set or hold-out will then contain some information derived from the characteristics of the other sets by virtue of the transformation derived from full-sample attributes being applied to each individual record. For critical applications where estimation of future performance is a priority, it is best to not pre-process data using transformations of this type when permitted by the algorithm.

Another consideration is that the "now shared" full-sample characteristics derived from the sample under consideration may vary greatly from that found in the actual population. When this occurs, pre-processing actions may not be able to be correctly and consistently applied to future observations. This can lead to even greater unexplained error rates when the model is deployed in production.

Another issue is that of the deployment requirements imposed by the pre-processing logic. Pre-processing logic becomes an implicit component of any model produced and must be identically reproducible (and successfully reproduced) when deploying a model against new data. As such, when a model is created, any operations performed upon the data between the acquisition of the original raw data and the submission of that data to a modeling algorithm for analysis is an integral part of the model logic. This means first of all, that in order for the model to be consistently and correctly applied, the pre-processing steps must be applied identically to all future data. This is not always possible depending on the pre-processing actions employed. It needs to be considered and understood that unless those actions can be applied identically and are available at the intended point of time for production use of the model, the model becomes a logically different model with a much greater degree of unknown behavior.

Related to this issue is the accompanying effect that observed values and ranges considered in the production of pre-processing parameters may not be correct with respect to the real population of records. This may lead to models which are unable to correctly predict some proportion of future data. When transformative pre-processing actions are undertaken that derive or rely upon sample aggregate measures, information specific to that sample are “baked” into all records in that data set. The resulting model is constructed with reliance upon this characteristic. As new data is evaluated, the sensitivity to this artifact varies between algorithms but the effect is always non-zero. The significance of this effect varies between algorithms and use cases, but when optimal predictive performance is paramount, this factor needs to be considered.

ADDITIONAL CONSIDERATIONS

Destruction of information: The activity of smoothing, averaging, and removal of outliers destroys some amount of information and detail present in the original data. It is analogous to the blurring of a photograph so that general trends become more apparent at the expense of precise detail. As the goal and purpose of predictive modeling is not to produce generalities (that is function of traditional statistical methods) but to classify or predict individual instances in as precise detail as possible, this practice has the potential to be counter productive to that goal.

There are also occasions where outliers represent attributes of the phenomenon under considerations. The elimination of such data points further removes meaningful information about the behavior to be understood and modeled. Outliers are information. They can be providing information about the subject under study or they can be providing information about the data collection process. There exist instances for both cases where that information is relevant and valuable in the effort of future prediction. Without a full understanding of the true underlying mechanics it is a challenging task to identify which information can be discarded to improve prediction capabilities. Misclassifying samples as discardable has the potential to destroy information critical to optimal prediction. For an illustration of the importance of maximizing behavior and representation in data sets and consequences of failing to do so, see the ZGP engine paper: "*Signal Detection Demonstration*".

It is also true that sometimes outliers do represent erroneous measurements. While it is common practice to presume that such outliers are a detriment to analysis, these outlier contain information that more sophisticated algorithms can exploit. For example, the presence of an outlier may be an indication that a collection process was conducted in a different manner and the record being assessed should be evaluated differently. An analysis engine can then take this into account and create a model that is more robust with regard to this artifact of the data collection process that in the future it might be called upon to interpret.

Further, what constitutes an “outlier” is often times an arbitrary decision made by a human participant in the modeling process. What constitutes a real departure from “reasonable” data is not a universal constant and depends upon data composition and use case. The essential consideration with respect to such values is whether they contribute to or detract from obtaining the best performing model. Computer algorithms have the potential to be much more effective at assessing when a value is useful for prediction and in which way it should be considered in a prediction. Removing this opportunity for a modeling engine only serves to inhibit the optimal modeling process.

ASSESSMENT OF PREDICTIVE MODELS BASED ON TRANFORMED DATA

One final note with regards to the estimation of model predictive capability related to the topic of hold-out and partition contamination mention previously. If a hold-out set or test partition is created from the original data for the purpose of estimating model performance and this estimation is intended to be used for model comparison or selection, it is not a valid use if compared against the metrics made by models build on data sets that were NOT constructed using transformative pre-processing actions. This is because evaluation against datasets that have had sample composition information imparted to all records will not provide the same degree of predictive assessment as models working with unprocessed data. To obtain a proper comparison between two such models, the comparison must be made against another data set collected and processed independently from the data set used to create the model.

CONSIDERATIONS WITH RESPECT TO MYDATAMODELS ZGP ENGINE

While some algorithms require comparative variables to be of similar magnitude and range or distribution, this is a characteristic that is only useful for a small range of mathematical relationships. In real data, the relations are generally much more diverse and complex.

The goal of such transformations is to homogenize the data such that elementary comparisons can be performed by simple analysis. With a mathematical expression engine such as that found in the ZGP engine, this is unnecessary. These simplistic transformations, as well as much more complex transformations, can be synthesized on demand as they are found to be predictively useful.

Additionally, when value transformations are found to be useful, the mathematical expression based modeling engine can construct transformations that are not dependent upon sample aggregate measures. Unlike models produced using transformed data, this leads to models which fully self-contain all logic required for model deployment in production.

Any gains that are observed when modeling transformatively pre-processed data can be the result of imparting information about the full sample composition. This can be due to models developed with some level of implicit optimization against all partitions constructed from the original sample set. It is not unusual for these apparent predictive performance increases to disappear when the models are evaluated with respect to new data which does not contain such information. In general models built without transformative pre-processing often behave more consistently than those built using data subjected to such transformations.

For the reasons described here as well as the absence of any value or correlation constraint in the ZGP engine and its ability to determine the merit and best use of outlier values, it is recommended that for optimal deployed results and capability, data not be subjected to transformative pre-processing when modeling with the ZGP engine.