Multivariate Metalog Distribution Model and Compression

Raul Rios

Lone Star Analysis

Compression is a key technology that enables efficient storage and transfer of large amounts of data. It is an operation that takes advantage of patterns in the raw underlying data representation and encodes it into a less redundant, more compact format. Probability distributions similarly reduce data by summarizing collections of individual datapoints into a smaller number of meta-parameters. A key aspect of data compression is the ability to recover the original data exactly (or near enough to be indistinguishable) by decoding the compressed data back into its native representation. Contrast this with data generated from a probability distribution fit which, statistically, never recover individual datapoints exactly, but instead capture aggregate behavior. In modeling and analysis scenarios, this is sufficient and desirable. In this paper, we demonstrate the uniquely flexible capability of the metalog distribution to fit data and produce an information-dense representation of that data.

Perhaps the most used probability distribution is the Gaussian distribution, in no small part due to its relevance to the central limit theorem. However, most real-world datasets are not truly Gaussian, which if paired with a Gaussian model fit, leads to a model with less analytic and predictive power. The task of choosing an appropriate distribution model to fit a dataset is often non-trivial. However, a recent innovation, the metalog distributions (Keelin, 2016), address these issues by proposing a distribution that is more flexible and easier to use for fitting datasets.

To examine the utility of the metalog distribution as a data summarization, or information compression tool, we use the Iris flower dataset – a common open-source benchmarking set used against classification / clustering solutions – as an example. We use 40 datapoints from each of three distinct flower classes: *setosa*, *versicolor*, and *virginica*. Each datapoint is a measurement of four features: petal length (PL), petal width (PW), sepal length (SL), and sepal width (SW). Since these features are not necessarily independent, we must estimate the covariance/correlation between features for each of the distinct classes. This information together with 4 independent metalog distribution fits – 1 for each feature – are sufficient for capturing the data of a single class. In this example we use 3rd-order metalogs, where the order refers to the number of fitting terms used in the metalog. Three terms are sufficient to capture the centration, shape, and skewness of a dataset.

To recreate data from the correlated multi-variate metalog model, we start by using statistical copulas (gaussian in this case) to generate correlated multi-variate CDF samples (Keelin, 2023). 1We then use the generated 4-D CDF samples, bound between 0 and 1, as input to the inverse-CDF function of the 4 metalog distributions to generate new synthetic data with the appropriate features and correlation. We repeat this process for each of the flower classes independently. In Figure 1Figure we can see that the synthetic points (right-half) match well to the original (left-half) groupings. It is useful to note that with the multivariate metalog model, we can generate as many synthetic datapoints as we like. This is a useful capability as many machine learning problems require significantly more data than is catalogued.

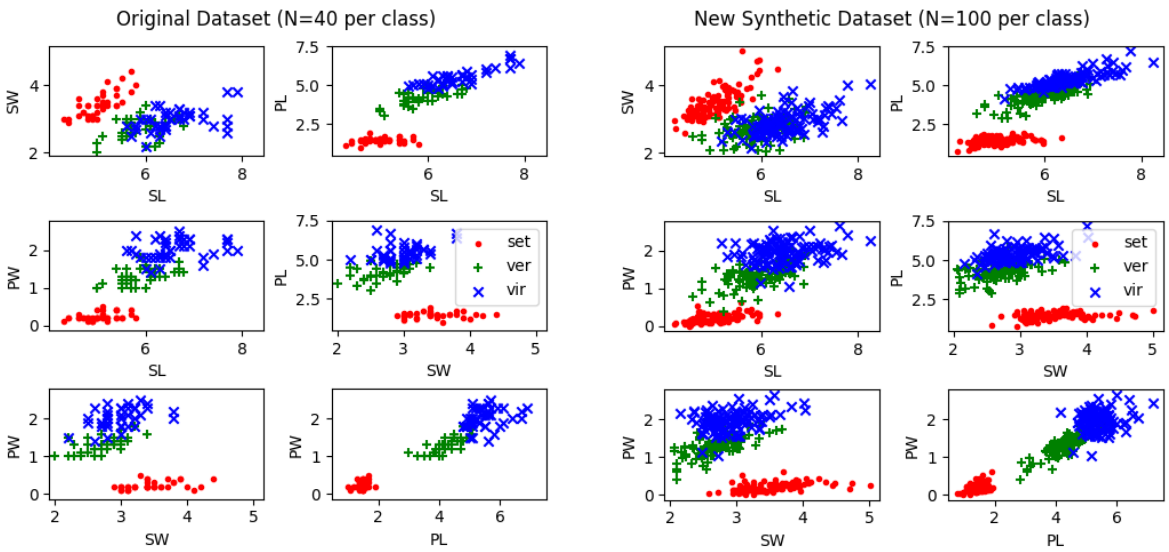

Figure 1. Comparison of Original and Synthetic Data for All Classes

As further confirmation that the synthetic dataset is representative of the original dataset, we trained a linear regression model on a training set of 30 datapoints per class of original data in one case, and an equal number of synthetically generated datapoints in another. Both models achieved similar class prediction accuracy (~93%) on a set of 20 test datapoints per class. This confirms the lossless transmission of information when used for machine learning solutions.

Returning to the analogy of distribution fits and compression, we examine the reduction from raw datapoints to model coefficients. In general, for a dataset with D dimensions, the covariance matrix contains $D(D+1)/2$ unique pairwise terms. Fitting D metalogs of order M generates DM coefficients. Thus, the total number of model parameters is $D(D+1)/2 + DM$. For the Iris example, D=4 and M=3 so we use 22 parameters for a single class, which means the effective "compression" with 40 initial datapoints is 0.55. Stated another way, we can take the complement of the compression ratio and say we achieved a data reduction of 45%. This data reduction percentage becomes larger (and better) with more data samples. Table 1 shows data reduction percentages for a range of M and D for a dataset of n=1000 samples. The data reductions scale as $1/n$.

Table 1. Metalog Data Reduction Percentage

		Dimension		
		4	10	30
Metalog Order	3	97.8%	91.5%	44.5%
	5	97.0%	89.5%	38.5%
	10	95.0%	84.5%	23.5%

Metalog distributions, much like other probability distributions, are useful tools for distilling information from a multitude of datapoints. What sets the metalog apart is the flexibility it offers to fit a wide variety of distributions without needing to choose a specific shape *a priori*. This enables the metalog to efficiently capture the nature of many real-world datasets. We have shown that a multivariate, correlated metalog model can faithfully reproduce realistic datapoints using a relatively small number of parameters. Since the model can generate a theoretically infinite amount of synthetic data, this feature is particularly useful for

augmenting sparse datasets that can then be used for training data-hungry neural networks. For these reasons, metalog probability distributions are a promising approach to distributed analytics processes that desire more economical data transmission.

References

Keelin, T. W. (2016, December). The Metalog Distributions. *Decision Analysis, 4*(4), 243-277.

Keelin, T. W. (2023, May 15). The Multivariate Metalog Distributions. Retrieved from The Metalog Distributions: http://metalogdistributions.com/publications.html