Statistical Similarity Evaluation Methodology

(Please find the next page for a process diagram of this methodology.)

For different needs, we proposed metrics evaluating the statistical similarity between the original data and synthetic data on a **column** level (pairwise columns) and **table** level (relationships between columns for each table).

To evaluate column-wise similarity, we propose using *KL-divergence* technique to calculate distribution similarity of the pair-columns. For more specific detailed statistics, *mean* and *variance* can be calculated for numerical columns. Probability density function can be used to visualize the distribution for an eyeballing comparison.

To evaluate table-wise similarity, we propose using *Auto-encoder* to first compress the original data to a certain dimensional latent representation. We will define the number of dimensions the latent representation based on whether we want to get a similarity score or visualization result of both tables. To get a similarity score, we will extract 1-dimensional latent representation for both tables and then apply similarity measurements, such as *Cosine Similarity* and *Euclidean distance*, to output the similarity score. To generate a visualization for eyeballing comparison of both tables, we will set a number for the dimension of the latent representation in auto-encoder that is higher than 2, and then use *PCA* or *t-SNE* to reduce the dimensions to only 2 for scatter plotting. Using *Autoencoder* to compress a high-dimensional dataset to a lower dimension makes using *t-SNE* for visualization feasible, since it requires low-dimensional input in terms of computation. In order to make sure both synthetic table and original table are transformed exactly in the same process, we will train the auto-encoder using the original dataset and then predict on the synthetic dataset using the model built. The same logic is applied for *PCA* and *t-SNE* transformation, where we use the same transformation function for both datasets.

Statistical Similarity Methodology

Column-wise Evaluation:

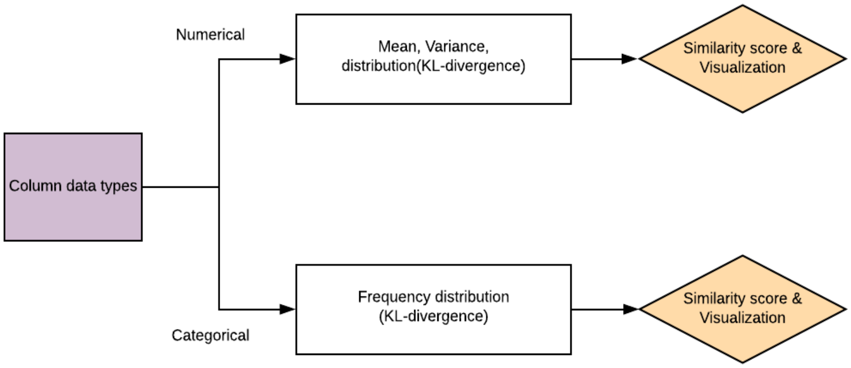
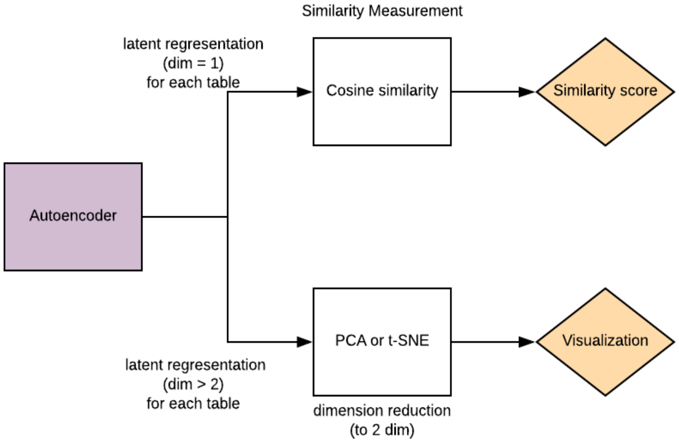

Numerical
Mean, Variance, distribution(KL-divergence)
Similarity score & Visualization
Column data types
Frequency distribution (KL-divergence)
Similarity score & Visualization
Categorical

Table-wise Evaluation:


Similarity Measurement
latent regresentation (dim = 1) for each table
Cosine similarity
Similarity score
Autoencoder
PCA or t-SNE
Visualization
latent regresentation (dim > 2) for each table
dimension reduction (to 2 dim)