MSBD6000B Project 1

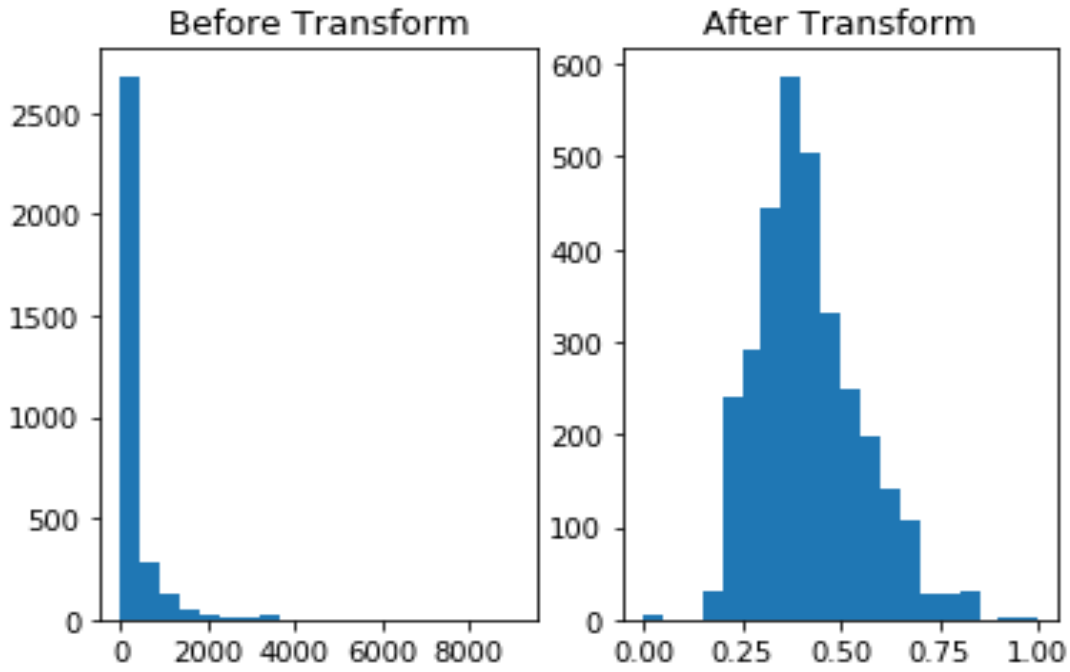
MAK, KIT HO
01652721

1. Data Preprocessing

Our first step is to get a summary of the data to have a sense of what it looks like. We see that different features span across very different ranges (max ranges from 2.17 for feature 46 to 9088 for feature 56), and that most of the data are zeros, so that almost all features are highly right-skewed.

We first standardize the data by 0-1 scaling to normalize the range of all features to 0-1, i.e. for each column X, transform it to (X – min(X)) / (max(X) – min(X)).

Next, to remedy the skewness of data, we take the 0.2 power of all data entrywise. The following picture shows the contrast of a sample feature (feature 56) before and after transform.



2. The Model

We used a SVM model (sklearn.svm.SVC) with linear kernel, and the penalty parameter is set to C=1.4.

The type of the model is determined by a (stratified) 80-20 train-test split, for which the performance of SVM with linear kernel is the best. The penalty parameter is then tuned using 5-fold cross validation (sklearn.model_selection.GridSearchCV) to the nearest 0.1.