# Understanding Data Virtualization for Learning Models

**Tal Ben Yakar**
tal-by@csail.mit.edu

## Abstract

Data are most crucial and essential building block for any data mining and AI application exist out there. More significantly, deep learning approaches require massive datasets. We know that the theory and algorithms have been around for quite a while however the ability to process the huge amounts of data brought us to the recent breakthroughs in the field. A challenge comes up in the case of a small dataset, compared to the required training data required. However, commonly, getting this data is neither an easy nor a cheap task, Many annotating services take advantage of the problem and charge for tagging data-sets campaigns, those could cost hundreds of dollars easily and still with an uncertain quality. Many wonder how to exploit the minimal data we have and still be able to learn well (generalize). In this paper, we overview methods for solving the tackle the problem and suggest solutions in order to overcome the challenge.

**Keywords**: AI, IoT, Data augmentation, virtualization , synthesized data, deep learning, cyber

## 1 Introduction

The task has gained many names, data generation, data augmentation, data synthesis...according to the McGraw-Hill , synthetic data are "any production data applicable to a given situation that are not obtained by direct measurement".

Synthesized data are used in a variety of fields. Cyber security for example, has many challenges , one of the largest is the lack of data and the smaller portion of data attacks, compared to the size of the recorded data. Therefore the data can also be used as a filter for information that otherwise change the results of the learning part and without it the performance would be very poor. Another consideration is the confidentiality of particular aspects of the data. Many times the particular aspects come about in the form of human information i.e. name, home address, IP address, telephone number, social security number, credit card number, etc.. Synthetic data are used in the process of learning, both for testing and training fraud detection systems, Synthetic data may seem as just a compilation of "made up" data, but there are specific algorithms and generators that are designed to create realistic data. Research related to medical applications and clinical trials or any other research such as medication consumptions, allergies data may generate data to aid in creating a baseline for future studies and testing. Another example is intrusion detection software, which are tested using synthetic data. The data are representative of the authentic data and may include intrusion instances that are not found in the authentic data. The synthetic data allows the software to recognize these situations and react accordingly. Synthetic data helps the models to generalize better, for instance the software trained only on authentic data, may not recognize another type of intrusion.

Many major AI breakthroughs have actually been constrained by the availability of high-quality training data sets, and not by algorithmic advances. Moreover, the preference of high-quality training data sets over purely algorithmic advances might allow an order-of-magnitude speedup in AI breakthroughs.

## 2 The problem

The problem at hand is to define what statistical estimators are required to generate synthetic data and how to generate large synthetic data from a small amount of real data, which might change per application. In this paper we solve this problem. The new generated data will be used for testing and validation tests for an IoT data engine.

# 3 Approaches

There are multiple ways to tackle the problem, the paper will discuss four effective ones :

- Methods for resampling
- Clustering
- Jittering
- Distribution estimation and sampling

# 4 Methods for resampling

## 4.1 Bootstrapping

Bootstrapping comes in handy when there is doubt that the usual distributional assumptions and asymptotic results are valid and accurate. Bootstrapping could be parametric or non-parametric. Generally bootstrapping follows the same basic steps:

1. Resample a given data set a specified number of times
2. Calculate a specific statistic from each sample.
3. Repeat steps (1) and (2) many times. The result will be a large collection of bootstrap replicate estimates for subsequent analysis.

**Sampling with/without replacement,** When we sample with replacement, the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second. Mathematically, this means that the covariance between the two is zero.

In sampling without replacement, the two sample values aren't independent. Practically, this means that what we got on the for the first one affects what we can get for the second one. Mathematically, this means that the covariance between the two isn't zero. That complicates the computations.

**Parametric vs Non-parametric.** Parametric bootstrap - a parametric model is fitted to the data, often by maximum likelihood, and samples of random numbers are drawn from this fitted model.

## 4.2 Monte carlo methods

Monte Carlo rely on repeated random sampling to obtain numerical results; typically one runs simulations many times over in order to obtain the distribution of an unknown probabilistic entity. The drawback is that the method requires heavy computation

## 4.3 Jackknife

The jackknife estimator of a parameter is found by systematically leaving out each observation from a dataset and calculating the estimate and then finding the average of these calculations. Given a sample of size N, the jackknife estimate is found by aggregating the estimates of each $N - 1$ estimate in the sample.

## 4.4 Cross-validation

Cross validation is perhaps most often known in the context of prediction, however, it can be used for resampling as well. The idea behind cross validation is that models should be tested with data that were not used to fit the model. Subsets of the data are held out for use as validation sets; a model is fit to the remaining data (a training set) and used to predict for the validation set. Averaging the quality of the predictions across the validation sets yields an overall measure of prediction accuracy.

One form of cross-validation leaves out a single observation at a time; this is similar to the jackknife. Unlike bootstrap and permutation tests the cross-validation dataset for training and testing is different.

## 4.5 Permutation tests

Permutation is a type of statistical significance test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points. Permutation tests exist for any test statistic, regardless of whether or not its distribution is known. Thus one is always free to choose the statistic which best discriminates between hypothesis and alternative and which minimizes losses."

The difference between permutation and bootstrap is that bootstraps sample with replacement, and permutations sample without replacement. The time order of the observations is lost and hence volatility clustering is lost — thus assuring that the samples are under the null hypothesis of no volatility clustering.

The permutations always have all of the same observations, so they are more like the original data than bootstrap samples. The expectation is that the permutation test should be more sensitive than a bootstrap test. The permutations destroy volatility clustering but do not add any other variability.

The major disadvantage is when there is seasonality or other patterns we might miss those. In addition, the time order of the observations is lost and hence volatility clustering is lost. Another issue is the variance compared to the original samples is small. Lastly, it's a good approach to evaluate statistics which is not the main goal here.

# 5 Clustering

Cluster analysis is used to 'learn' the number and characteristics of the components of the mixture distribution. For this purpose, similar elements of the sample are assigned to groups (clusters). For the purpose discussed here, we can fit a clustering model to discover the shape of the data. Selecting the number of clusters and other parameters using a clustering metric. Then, generate data using random samples from clusters.

Two approaches for first stage clustering could be GMM and K-means, we discuss one of them here

## 5.1 GMM

A Gaussian Mixture Model (GMM) is a parametric probability density function represented as a weighted sum of Gaussian component densities. GMMs are commonly used as a parametric model of the probability distribution of continuous measurements or features in a biometric system, such as vocal-tract related spectral features in a speaker recognition system. GMM parameters are estimated from training data using the iterative Expectation-Maximization (EM) algorithm or Maximum A Posteriori (MAP) estimation from a well-trained prior model. Ideally, a cluster represents all of the elements drawn from one population of the mixture.

# 6 Jittering

Real data samples contain random measurement errors. Even if the same objects were observed multiple times under the same experimental conditions, the data are likely to be different. These differences can be simulated by generating copies of the original sample and adding random values to each of these data sets. The normal distribution with zero mean is traditionally used for this purpose. If estimates of the measurement error exist, this information can be utilized to define the parameters of the error distribution.

In other words, generate a new data ie take the existing set and add white noise - advantage - if there are patterns in the data we will still keep them.

Drawbacks

- The data are assumed to be normally distributed
- Underlying assumption is that the features are independent which is hard one and might not simulate the case.

# 7 Distribution estimation and sampling

Estimate the distribution of the data and sample within the distribution parameters. For example use a test to fit the data to normal distribution and then sample by the estimated mean and variance.

Drawbacks

- If the distribution prediction is wrong it might lead to a bad results.
- There is no well known method to do so.

## 8 Conclusion

Bootstrapping (drawing with replacement) is perhaps the most widely known and recommended resampling approach, because it is a standard approach for statistical inference methods (Efron & Tibshirani, 1993). If the sample size is large and the true distribution is well represented by the data, bootstrapping may also be useful for the validation of clustering results. That is, other resampling schemes may not lead to more accurate results (cf. Minaei-Bidgoli et al., 2004). Under these circumstances the user may prefer bootstrapping, because no control parameter has to be set.

All other methods are not well known and some intuitively work for this usecase. One need to choose the method with consideration of the data in hand.

## References


[1] Chihara, Laura M., and Tim C. Hesterberg. *Mathematical statistics with resampling and R*. John Wiley & Sons, 2012.

[2] Horvitz, Daniel G., and Donovan J. Thompson. "A generalization of sampling without replacement from a finite universe." *Journal of the American statistical Association* 47.260 (1952): 663-685.

[3] Schluter, Dolph. "Estimating the form of natural selection on a quantitative trait." *Evolution* (1988): 849-861.

[4] Tseng, George C., and Wing H. Wong. "Tight Clustering: A Resampling‐Based Approach for Identifying Stable and Tight Patterns in Data." *Biometrics* 61.1 (2005): 10-16.

[5] Hastie, Trevor, and Robert Tibshirani. "Discriminant analysis by Gaussian mixtures." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 155-176.

[6] Reynolds, Douglas. "Gaussian mixture models." *Encyclopedia of biometrics* (2015): 827-832.