Arthur Lapin

Statistical Bootstrapping

The objective of this document is to explain the mechanics behind bootstrapping and provide some insight as to why this statistical technique was used.

An experiment or test is usually started with stating the hypothesis or initial idea. The next step is to perform some type of test or tests that provide results which either help prove the hypothesis true or prove the hypothesis false. These tests vary in many aspects such as context, size, domain, rate, etc. One common test involves gathering a very large sample set and finding a general correlation in that data set. As the sample size increases the correlation is easier to find and provides stronger results for proving the hypothesis true or false. The reason being is that the individual gathering the sample set is trying to build the best model that mimics the actual population model that this individual is interested in but cannot gather because it is impossible. For example, an individual is interested in finding the distribution of the height of human males in the world. It is obvious that this individual cannot measure the height of all human males in the world. Thus the individual results to measuring as many human males as he possibly can and use that as a sort of surrogate population model.

Bootstrapping is a technique that uses this idea but takes it to the next level. The objective of the bootstrapping technique is of course to find the population model. As it cannot be seen and impossible to gather, a large sample set is gathered and this sample set is then called the surrogate model. Bootstrapping then tries to find information about the actual population model from analyzing the surrogate model and finding an accurate correlation between the surrogate population model and the actual population model. The analysis of the surrogate population distribution is done by creating an N number of bootstrap distributions by randomly sampling data with replacement from the surrogate model. However, for the bootstrapping technique to actually work all the bootstrapping distributions must be the same size as the surrogate distribution (same sample size). Because all the bootstrapping distributions were created with random sampling and replacement, all the distributions are different but contain some correlation between each other and to the surrogate distribution.

After the N number of bootstrapping distributions are created, the mean and variance of every distribution is calculated. All the means and variances are then combined and two normal distributions are created, one normal distribution for all the means and one normal distribution for all the variances. Normal distributions are usually bell shaped and the mean and variance values at the center of their respective normal distributions are the most likely values for the mean and variance. Knowing mean and variance of any distribution allows the parameters of well-known distribution to be calculated. For example, knowing the mean and variance of the

gamma distribution allows the gamma parameters alpha and theta to be calculated from the equations mean = alpha * theta and variance = alpha * (theta)^2.

By finding the mean and variance of the surrogate population distribution, its parameters can also be calculated. Then, knowing the most likely (top of bell shaped normal distribution) bootstrapping parameters and the surrogate population parameters, the difference of these parameters can be calculated. As the bootstrapping distributions are sample distributions of the surrogate population distribution, the surrogate population distribution is itself a sample distribution of the actual population distribution. This information then also provides information that the difference between the bootstrap distribution and surrogate model distribution parameters is equal to the difference between the surrogate distribution and the actual population distribution. This correlation then allows for the parameters of the actual population distribution to be calculated and knowing the values of these parameters knowledge can be gained about the population distribution.