Project Proposal

Fabian Juarez, Surya Sama

April 2022

Introduction

The Central Limit Theorem in probability theory states that for X_1, X_2, X_3, ..., X_n,.. random samples which are independent and identically distributed random variables drawn from a distribution of expected value given by μ and a finite variance given by σ^2, as n gets larger, the distribution of the difference between the sample average $\bar{X}_n$ and its limit μ, when multiplied by the factor $\sqrt{n}$,

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{a} \mathcal{N}(0, \sigma^2)$$

approximates the **Normal Distribution** with mean 0 and variance σ^2. For large enough n, the distribution of $\bar{X}_n$ is close to the normal distribution with mean μ and variance $\frac{\sigma^2}{n}$.

We want to research if the central limit theorem applies to the different performance metrics used across a variety of machine learning algorithms.

Approach

Let $s \in \mathbf{S}$ such that $s \in (0, 1)$ be a set of proportions of training data for a given model. And let N be the number of iterations.

For each s, we plan on using boot-strapping to create training sets of size s and then run the machine learning algorithm to fit the model to the given training sample and then compute the performance metrics on both training data and the test data. We will iterate this procedure N times and collect all the performance metrics in their respective vectors $\mathbf{P}^s_m$, m being the metric and s being the size of the training sample.

Once we have the test and training metrics, histograms of the sorted data would be plotted to visually identify if the data is distributed according to any probability distribution. We want to plot a QQ plot to see how far the data is

distributed from the Normal distribution. Also we would like to perform Jarque-Bera test to check the null hypothesis that the data is distributed normally. We also plan on checking if the data follows any other distributions like the Cauchy distribution and any other possible distribution.

Now, we repeat this analysis for the remaining $s \in \mathbf{S}$ and observe whether the distributions of the performance metrics are converging to any of the above mentioned distributions as s increases.

We want to try using the above methodology to try and determine the standard errors of the coefficients of Ordinary least squares regression and confirm it with the actual OLS results to verify our methodology.

Models & Data

We want to try this approach on a set of machine learning classifiers, neural network classifiers, and regressor algorithms. Since the performance metrics for the regressor algorithms are usually, the Mean squared error, Mean absolute error, we conjecture that those metrics might converge to normal distribution since the central limit theorem is valid for sample mean. We want to explore how the metrics for classification algorithms behave.

We are planning on using different kinds of models available on Kaggle and on tutorials for tensorflow to perform this experiment on. We are doing this because, the models available on kaggle and tensorflow are well behaved and are apt for the data sets being used in the model and the kind of problem that is being modeled. The data we are going to use is also the data being used by the models in kaggle and tensorflow.

Further research

We also want to examine the behaviour of the performance metrics from the models that are not suitable for the data sets as well and check if there is any distinction between the way the sampling distributions of the metrics converge or diverge.

We want to explore what happens if the number of iterations in bootstrapping process $\mathbf{N}$ also keeps on increasing along with the sample size s of the training data.