Finding Empirical Data Distribution

Sapir Rubin, 301659751, rubinsapir@gmail.com

February 2022

Abstract

Probability distributions are a fundamental concept in statistics. They are used both on a theoretical level and a practical level. Unfortunately, the detection and characterization of empirical distribution might be complicated when just part of the data is distributed according to known distribution while the rest are outliers. Here we will present a statistical framework for detecting and quantifying hidden distribution in empirical data. Our approach combines statistical methods (as shown in class) with Genetic Algorithm and taking into consideration each time part of the data. we will evaluate the effectiveness of the approach with real data and give critical comparisons to previous approach. We will see that in most cases we will find that the approach will managed to find the hidden distribution with high probability.

1 Problem description

Finding empirical data distribution is important part within the Machine Learning pipeline. Understanding the data behaviour can help us with :

- To calculate confidence intervals for parameters and to calculate critical regions for hypothesis tests.
- In the case of univariate data, it is often used to determine a reasonable distributional model for the data.
- Statistical intervals and hypothesis tests are often based on specific distributional assumptions.
- Continuous probability distributions are often used in machine learning models, most notably in the distribution of numerical input and output variables for models and in the distribution of errors made by models.

Current methods, as shown in class, are searching for a match in the entire given data. In fact, if a given data has group of outliers it will eventually let to all entire-data methods to fail.

2 Solution overview

The solution combines maximum-likelihood estimation (MLE) methods with goodness-of-fit tests based on minimizing the Kolmogorov-Smirnov (KS) statistic. Optimization is performed using a genetic algorithm (GA). Within this approach, we will

1. Use MLE to estimate (based on the empirical data) the distribution's parameters (as shown in class).

2. Use goodness-of-fit tests based on minimizing the Kolmogorov-Smirnov (KS) statistic (as shown in class).

   Assume $\mathcal{S}$ is the set of the empirical data, $n = |\mathcal{S}|$, and $\mathcal{S}' \subset \mathcal{S}$. Define $i : \mathcal{S}' \to \mathcal{N}$ where $i(x_j)$ is the index of x_j in $\mathcal{S}'$. Define S^* to be the empirical CDF of the data for the observations only in $\mathcal{S}'$ Hence,

   $$S^*(x) = \frac{i(x)}{n}. \tag{1}$$

   In addition, we will use $P(x)$ which is CDF for the distribution in question that best fits the data and will aim to find $\mathcal{S}'$ which minimize KS statistic:

   $$KS(\mathcal{S}') = \max_{x \in \mathcal{S}'} |S^*(x) - P(x)|. \tag{2}$$

3. Use a genetic algorithm to find the largest subset of data from the empirical one that applies to the distribution in question.

4. Use again goodness-of-fit test for analyzing performance.

A Genetic Algorithm (GA) is an optimization technique that mimics biological evolution as a problem-solving strategy. It is based on Darwinian principles of evolution to optimize a population of candidate solutions towards a predefined fitness.

GA uses a chromosome-like data structure and evolve the chromosomes using selection, recombination and mutation operators. The process usually begins with randomly generated population of chromosomes, which represent all possible solutions of a problem. From each chromosome, different positions are encoded as bits, characters or numbers. These positions could be referred to as genes. An evaluation function is used to calculate the goodness of each chromosome according to the desired solution; this function is known as the "Fitness Function". During the process of evaluation "Crossover" is used to simulate natural reproduction and "Mutation" is used to mutation of species. Mutation occurs to maintain diversity within the population and prevent premature convergence. See Fig. [1] for a schematic flowchart of a standard GA. In our case, we would like to find subset of the empirical data that resembles a power-law distribution. GA will find the best chromosomes (data points) that built the fittest (minimizing the optimization problem Eq.(2)) population (subset). The GA's principles in our implementation are:

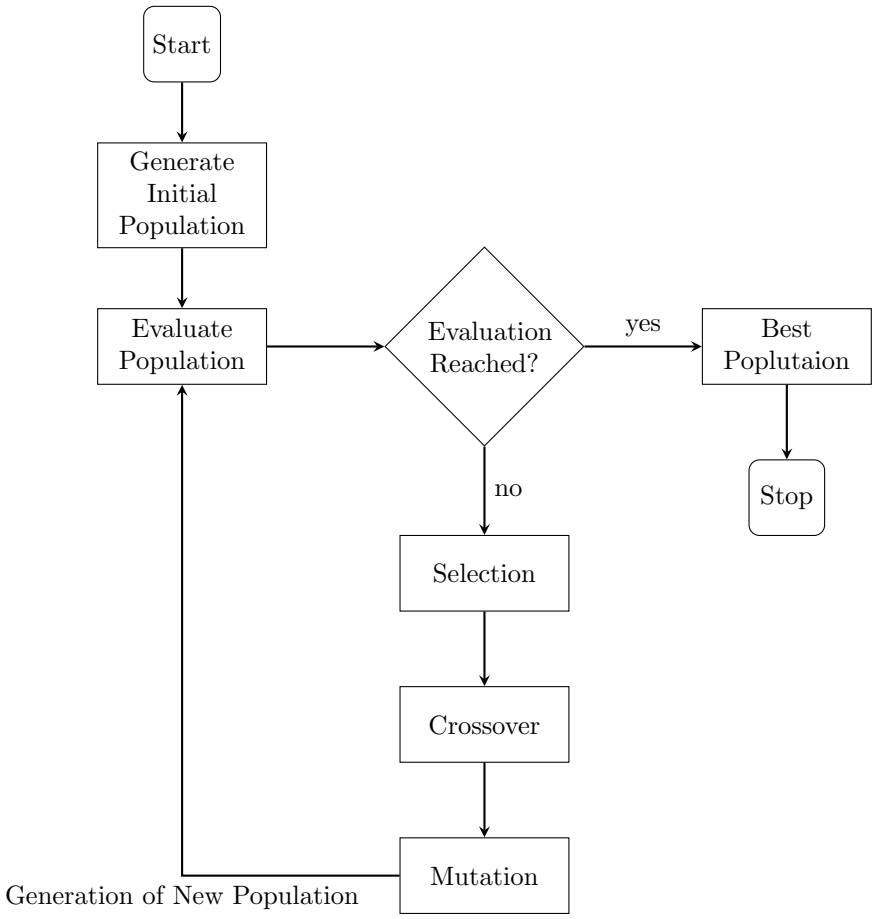

Figure 1: Flowchart of a standard Genetic Algorithm

- Selection - In each step, based on the fitness function, we are going to select the best individuals within the current population as *parents* for mating. We will take the best k parents where k is hyper-parameter, these parents holds the top k scores from the entire population.
- Crossover - Our crossover approach implements crossover using a draw of a random number in the range [0,1] to determine if crossover is performed. If so this involves selecting a random split point on the bit string, then creating a offspring with the bits up to the split point from the first parent and from the split point to the end of the string from the second parent.
- Mutation - Mutation involves shutting bits - $(bit = 0)$ - in created offspring candidate solutions. Each bit in a binary-valued chromosome typically has a small probability of being flipped. For a chromosome with m bits, this mutation rate is typically set to $1/m$, yielding an average of one mutation

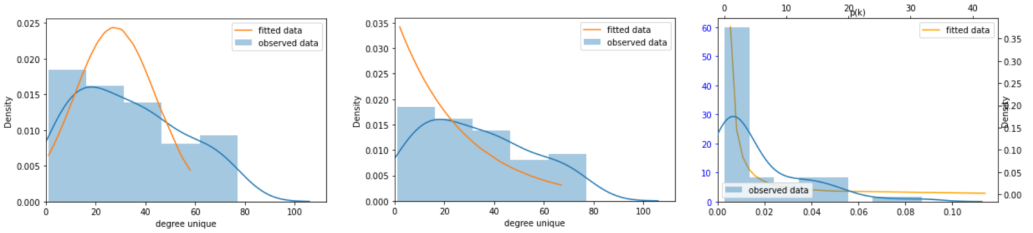

Figure 2: Facebook social network graph fits with (a) Normal Distribution (b) Exponential Distribution (c) Power-Law Distribution. The empirical data histograms and PDF in blow and the fitted PDF in orange. We can see that when removing the correct samples each of the distribution can fit the empirical data.

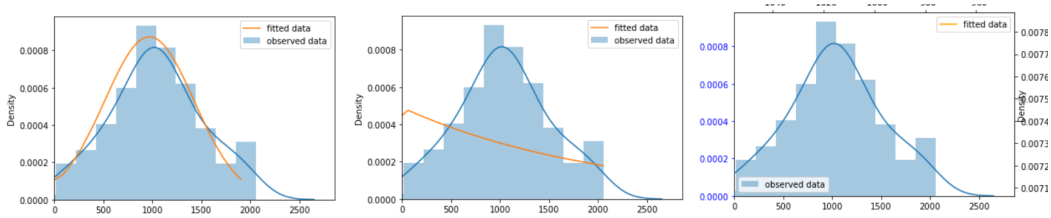

Figure 3: Company "ABC" profit fits with (a) Normal Distribution (b) Exponential Distribution (c) Power-Law Distribution. The empirical data histograms and PDF in blow and the fitted PDF in orange. In (a) it seems that the suggested methods removed all non-normal samples.

per offspring chromosome.

3 Experimental evaluation

In this section, as a demonstration of the utility of the methods described above, we apply them to a real world data sets representing measurements of quantities. As we will see, the results indicate that removing the correct part of the data might achieve better fit.

The three data sets studied are drawn from a different branches of human endeavor, including social sciences, real estate, economy. They are as follows:

- Distribution degree of Facebook social network graph
- Company "ABC" profit (normal distributed samples with 2% of manually added outliers)
- Boston house prices

All above data sets were examined to fit any of below, with the methods described above or with the methods presented in class (i.e. with no change):

- Normal distribution
- Power-Law distribution
- Exponential distribution

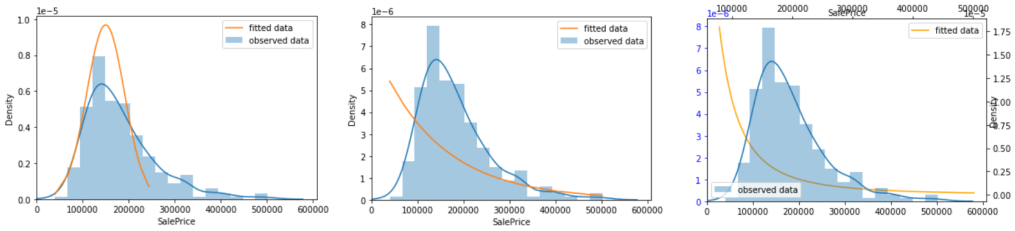

Figure 4: Boston house prices fits with (a) Normal Distribution (b) Exponential Distribution (c) Power-Law Distribution. The empirical data histograms and PDF in blow and the fitted PDF in orange. Suggested method manage to find solid fit in (a) and bad fits in (b) and (c)

Suggested Method - Power Law Distribution			
Parameters	Facebook	House Prices	Profit
α	1.377	2.41	-6.47
KS statisic	0.31	0.26	NaN
p-value	0.971	0	NaN
Existed Method - Power Law Distribution			
α	1.258	1.69	-8.05
KS statisic	0.084	0.192	NaN
p-value	0.779	0	NaN

Table 1: Methods comparison for examine Power-Law distribution fit on three data sets

Figs [2], [3] and [4] shows the suggested method fit to the empirical data. and tables [1], [2] and [3] detailed the experiments results with comparison of suggested method and current methods (which showed in class).

The most impressive results is fitting Normal distribution to "Company "ABC" profit" and "Boston house prices" dataset, removing the correct part from the data ended with improving *p-value* from 0.17 to 0.993 and from 0.0013 to 0.98 respectively, In addition we can observe in Fig. [3] and [4] that the resulted fit's PDF emphasise the bell curve in the right places.

We can see that for each distributions fit to Facebook dataset can be a good fit the data after suggested method is applied.

For House Prices dataset we can observe that Exponential and Power-Law distribution can not fit the data, but Normal distribution might be a very good one, depending which and how many samples are removed.

Suggested Method - Normal Distribution			
Parameters	Facebook	House Prices	Profit
μ	27.7	149891.34	969.96
σ	16.36	41210.97	457.09
KS statisic	0.081	0.039	0.04
p-value	0.97	0.98	0.993
Existed Method - Normal Distribution			
μ	33.0	181451.52	972.67
σ	21.48	76050.92	573.07
KS statisic	0.083	0.12	0.076
p-value	0.786	0.0013	0.17

Table 2: Methods comparison for examine Normal distribution fit on three data sets

Suggested Method - Exponential Distribution			
Parameters	Facebook	House Prices	Profit
λ	0.036	~ 0	0.0004
KS statisic	0.15	0.44	0.369
p-value	0.41	0	0
Existed Method - Exponential Distribution			
λ	0.031	~ 0	0.0005
KS statisic	0.96	0.99	0.99
p-value	0	0	0

Table 3: Methods comparison for examine Exponential distribution fit on three data sets

4 Related work

As seen in section above, there are many methods for examining if entire dataset is fitting according to a known theoretical distributions. I got inspired by particular paper - *"Power-law distributions in empirical data."* by Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. The method proposed in this paper mainly dealing with finding the lower bounder for Power-Law fit in given empirical data and not searching in the entire dataset for cases where the data is spread all over the support of the distribution. In addition, commonly used graphic methods for analyzing empirical data, such as least-squares fitting, can produce substantially inaccurate estimates of parameters for the distribution in question, and even in cases where such methods return accurate answers they are still unsatisfactory because they give no indication of whether the data obey a the distribution at all.

5 Conclusions

We have seen a method that combine statistical models with Genetic Algorithm for searching for the best subset in a given dataset that fit to some theoretical distribution. We saw cases there is no improvement from other existed methods,

cases where there is a slight improvement and cases where the suggested method finds an accurate subset for a theoretical distribution fit. I have learned from this project that with smart use in existing tools you can improve your own project dramatically. In addition the option for analyzing and learning on each theoretical distribution (not only the ones mentioned in this work) will always bring some new insights that help build our self as mathematical-statistical-computer scientists.