Assignment 1

CSEE5590/490: AI CyberSecurity Programming (2019 Summer)

Byron Wheeler

June 23, 2019

# Introduction

The purpose of Assignment 1 is to rebalance the randomized response done in In-Class Programming 3(ICP3) on 6/13/2019. The changes will include adding a noise parameter to the query. This noise parameter will be a percentage of the noise given to the randomization process. For ICP3, the noise was set to 50%. After the randomized response is rebalanced, the query will be performed with the mean function on datasets that range from 10, 100, and 1000. For each dataset the true query and the “noisy” query will be recorded.

# Objectives

* Rebalance the Randomized response
* Record running the query multiple times, varying the noise in steps of 0.1, starting with a value of 0.1 and ending with a value of 0.9.
* Perform a comparison on varying the noise and the size of the dataset.

# Approach/Methods

The work of this assignment will be done in the ICP3 Jupyter Notebook. The Google Colab research website will be used to implement the notebook.

# Workflow

1. Import required libraries
2. Define function for creating a database of a specific size. The data in the database doesn’t matter for this example, so it’ll be random binary values(0,1).
3. Define a function for flipping a coin for each element in a database
4. Define a function for querying the database, and using Randomized Response to return mean of the data with and without noise.
5. Iterate through all the defined database sizes, and noise percentages. In each iteration, calculate the query on the original, and noisy data. Compare the difference.

# Datasets

The datasets used were Radom binary values(0,1). They varied in length from 10, 100, and 1000. Technically speaking the noisy data is a dataset, but it is really the same as the original random one. It just has had noise added to it.

# Parameters

* Database sizes: 10, 100, 1000, 10000, 100000
* Noise percentage(% Heads vs Tails): 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%

# Evaluation and Discussion

According to the output, increasing the size of the database decreases the difference between the original dataset and the augmented dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Database Size | Noise | Original DB | Augmented DB | Difference |
| 10 | 0.1 | 0.5 | 0.5 | 0 |
| 10 | 0.2 | 0.5 | 0.3 | 0.2 |
| 10 | 0.3 | 0.4 | 0.4 | 0 |
| 10 | 0.4 | 0.5 | 0.7 | 0.2 |
| 10 | 0.5 | 0.3 | 0.5 | 0.2 |
| 10 | 0.6 | 0.4 | 0.8 | 0.4 |
| 10 | 0.7 | 0.6 | 0.5 | 0.1 |
| 10 | 0.8 | 0.4 | 0.7 | 0.3 |
| 10 | 0.9 | 0.7 | 0.7 | 0 |
| 100 | 0.1 | 0.5 | 0.48 | 0.02 |
| 100 | 0.2 | 0.49 | 0.47 | 0.02 |
| 100 | 0.3 | 0.5 | 0.52 | 0.02 |
| 100 | 0.4 | 0.5 | 0.54 | 0.04 |
| 100 | 0.5 | 0.43 | 0.5 | 0.07 |
| 100 | 0.6 | 0.42 | 0.51 | 0.09 |
| 100 | 0.7 | 0.51 | 0.54 | 0.03 |
| 100 | 0.8 | 0.54 | 0.52 | 0.02 |
| 100 | 0.9 | 0.5 | 0.49 | 0.01 |
| 1000 | 0.1 | 0.498 | 0.492 | 0.006 |
| 1000 | 0.2 | 0.484 | 0.496 | 0.012 |
| 1000 | 0.3 | 0.504 | 0.487 | 0.017 |
| 1000 | 0.4 | 0.51 | 0.488 | 0.022 |
| 1000 | 0.5 | 0.488 | 0.49 | 0.002 |
| 1000 | 0.6 | 0.52 | 0.529 | 0.009 |
| 1000 | 0.7 | 0.49 | 0.486 | 0.004 |
| 1000 | 0.8 | 0.511 | 0.502 | 0.009 |
| 1000 | 0.9 | 0.499 | 0.479 | 0.02 |
| 10000 | 0.1 | 0.5038 | 0.5029 | 0.0009 |
| 10000 | 0.2 | 0.4971 | 0.4945 | 0.0026 |
| 10000 | 0.3 | 0.4975 | 0.5015 | 0.004 |
| 10000 | 0.4 | 0.5 | 0.5012 | 0.0012 |
| 10000 | 0.5 | 0.5002 | 0.4998 | 0.0004 |
| 10000 | 0.6 | 0.5042 | 0.4944 | 0.0098 |
| 10000 | 0.7 | 0.5055 | 0.5053 | 0.0002 |
| 10000 | 0.8 | 0.4961 | 0.4969 | 0.0008 |
| 10000 | 0.9 | 0.5045 | 0.4947 | 0.0098 |
| 100000 | 0.1 | 0.50067 | 0.50064 | 3.00E-05 |
| 100000 | 0.2 | 0.50305 | 0.50301 | 4.01E-05 |
| 100000 | 0.3 | 0.50124 | 0.5007 | 0.00054 |
| 100000 | 0.4 | 0.49927 | 0.49952 | 0.00025 |
| 100000 | 0.5 | 0.49812 | 0.50076 | 0.00264 |
| 100000 | 0.6 | 0.50083 | 0.50101 | 0.00018 |
| 100000 | 0.7 | 0.50224 | 0.50099 | 0.00125 |
| 100000 | 0.8 | 0.49821 | 0.50051 | 0.0023 |
| 100000 | 0.9 | 0.49852 | 0.50084 | 0.00232 |

# Conclusion

Tweaking the “noise”, making the data lean to heads or tails, didn’t change the output of the data. But, it was more useful to analysts in that they can get a higher probability of correctness when they aggregate the dataset(noisy).