Sahil Punchhi, z5204256 COMP 9417, 2019T2

**Laboratory Earthquake Prediction from real-time Seismic data**

COMP 9417 Machine Learning Project

Introduction

Forecasting earthquakes is one of the most significant problems in Earth science because of their devastating consequences. Current scientific studies related to earthquake forecasting focus on three key points: when the event will occur, where it will occur, and how large it will be. In this project, an attempt has been made to address when the earthquake will take place. Specifically, the time remaining before laboratory earthquakes occur from real-time seismic data is predicted. The seismic signal originates from continuous grain motions of the fault gouge as the fault blocks displace. The laboratory test applies shear forces to a sample of earth and rock containing a fault line. If the physics are ultimately shown to scale from the laboratory to the field, researchers will have the potential to improve earthquake hazard assessments that could save lives and billions of dollars in infrastructure. The topic selected for the project is chosen from Kaggle competitions where  [Los Alamos National Laboratory](https://www.lanl.gov/) (LANL) team has provided an interesting and a challenging dataset with notably more aperiodic earthquake failures and the objective is to predict the failures of each test data set.

A close up of an animal

Description automatically generated

Related Work

A lot of work has been going on at Los Alamos National Laboratory to study the physics which drives the geological faults which could help in predicting earthquakes more accurately. The collapse of stress chains inside the earthquake gouge gives rise to seismic signals in the lab which directs the study to the similar phenomenon taking place inside the Earth. This work was published in Physical Review Letters by Ke Gao, a computational geophysicist in the Geophysics group at LANL under the heading “From Stress Chains to Acoustic Emission”. Stresses are transmitted from one side of fault block to the other in bridges composed of grains called stress chains. According to recent developments, machine learning algorithms using simple decision trees have given a lot of insightful information to the researchers at LANL about the principles of physics in the mechanics of earthquakes which even deep neural network failed to achieve. Statistics around the seismic data from the laboratory experiments answers various questions established by decision trees and the algorithm branches to a new decision based on the previous decision, thus forming a tree with branches.

Bertrand Rouet-Leduc investigates that machine learning discerns the frictional state when applied to laboratory seismic data

recorded during a shear experiment in his geophysical research letter “Machine Learning Predicts Laboratory Earthquakes” *(*

[*https://doi.org/10.1002/2017GL074677*](https://doi.org/10.1002/2017GL074677)*).* He further mentions that working through continuous seismic data would help in advancing through the identification of currently unknown signals, in deeper understanding of earthquake fault physics and in the accuracy of fault failure times. Paul Johnson from LANL has applied machine learning approach to study the acoustic signal data from slow displacement of formerly adjacent points on opposite sides of a fault, measured on the fault surface, in the real-world scenario from earthquake prone regions in America and New Zealand.

Experimental Data

The goal of this project activity is to use seismic signals to predict the timing of laboratory earthquakes. The data comes from a well-known experimental setup used to study earthquake physics at Los Alamos National Laboratory. The seismic input signal (acoustic\_data) is used to predict the time remaining before the next laboratory earthquake (time\_to\_failure). The training data is a single, continuous segment of experimental data consisting of 629,145,480 rows each containing the seismic signal and the corresponding time to failure.

The test data consists of a folder containing many small segments. The data within each test file is continuous, but the test files do not represent a continuous segment of the experiment; thus, the predictions cannot be assumed to follow the same regular pattern seen in the training file.

Kaggle category: Research Prediction

Kaggle competition Link: <https://www.kaggle.com/c/LANL-Earthquake-Prediction>

Data Exploration

The data set given has a very large dimension, more than 600 million rows of data. Descriptive summaries are obtained for both the columns, acoustic\_data and time\_to\_failure which are presented below.

|  |  |
| --- | --- |
| acoustic\_data | |
| count | 629,145,500 |
| mean | 4.52 |
| std | 10.74 |
| min | -5,515 |
| max | 5,444 |

|  |  |
| --- | --- |
| time\_to\_failure | |
| count | 629,145,500 |
| mean | 5.68 |
| std | 3.67 |
| min | 0.00 |
| max | 16.11 |

Acoustic data distribution is plotted for 0.5% of the overall data and outliers are observed in both the directions. Then, only the values which are between -20 and 20 are considered and the distribution is again plotted using distplot() function from seaborn library. This function draws a histogram and fits a kernel density estimate (KDE) for a univariate distribution. The black line is the closest normal distribution (gaussian) possible.

A picture containing text

Description automatically generated

Time to failure (given in seconds) is the target variable and below is the univariate distribution obtained for it.

A close up of a map

Description automatically generated

In the next graph, both the variables are simultaneously plotted to see how both the variables change over time. Initially 0.5% of the overall data is taken by sampling every 200 points of the data. The acoustic data reveals complex oscillations with varying amplitude, but just before each failure there is a significant increase in the amplitude of the seismic signal. The graph also shows that amplitude also increases somewhere midway between two consecutive failure points. The time to failure decreases slowly linearly before each failure.

A screenshot of a cell phone

Description automatically generated

Additionally, after looking at the above graph, more analysis is done by closely observing first 2% of the data. Carefully looking at the graph now, it is observed that large amplitude of the seismic signal is not just before the failure time. Also, there are repetitive oscillations with comparatively higher amplitude after this large one.

A screenshot of a social media post

Description automatically generated

When 100% of the data is plotted, it is observed that there are 16 earthquakes in the training data.

Acoustic values greater than 450 are considered high and more than 80% of high acoustic values are around 0.3 seconds before an earthquake, which is an important finding.

Feature Engineering

References