LANL Earthquake Prediction

https://www.kaggle.com/c/LANL-Earthquake-Prediction

This document describes the steps involved to solve the given problem. Follow this document and .ipynb document simultaneously to completely understand the steps mentioned below. We will start by exploring the data, followed by feature extraction, and finally, use an appropriate model to make predictions after training and validation. If any further clarification is required regarding the code, write me at sunny@doeity.com.

Sequence of Steps:

1. Load the training data from the given dataset

This includes loading and reading data using Pandas.

2. Visualise the training data - Input acoustic_signal and time_to_failure

Since the dataset is massive, we plot only 1 in every 100 points that is given to us, or in other words 1% of the given data. The data is recorded in segments at a frequency of 4MHz, each segment lasting for 0.0375 seconds, thus constituting 150,000 data points.

We see that the acoustic signal shows a massive spike, followed by a quieter than usual period in the build up to an Earthquake, when time_to_failure hits 0.

3. Extract features using appropriate mathematical functions

The entire data is first split into segments of 150,000 data points each, giving us 4194 segments to train on. In this step, the data to be trained on is created by extracting features. To capture the inherent frequency-magnitude characteristics of the wave, we use the Fast Fourier Transform (FFT) on each segment and take the real and imaginary components and calculate the mean, standard deviation, maximum and minimum for each.

In addition, we take several other features such as the quartiles, Median Absolute Deviation (MAD), in addition to basic statistical features of standard deviation, mean, max and min. Some other functions such as kurtosis and skew are also used.

4. Train and Validate using an appropriate model

We start off by creating the training data, and targets. This is followed by scaling the training data, and the same is then done for the testing data.

5. Make predictions

The model is chosen based on the cross validation results. It can be appropriately chosen as per one's own intuition and reasoning.