

Predicting Earthquakes

Exposé for the “Deep Learning: Architectures and Methods” project

David Gengenbach (info@davidgengenbach.de)

Tom König (TODO)



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Informatik

Abstract

For our “Deep Learning: Artitectures and Methods” project we participate in an online machine learning challenge called **LANL-Earthquake-Prediction** hosted on the Kaggle platform. Here, we try to predict when the next earthquake is happening using historical data provided by the **Los Almos National Labratory**.

We hope that the community-driven design of Kaggle can result in a great learning opportunity. Another big motivator is the open-source character of the Kaggle community which in turn enables the comparison and sharing of different approaches or solutions.

1 The challenge

“Forecasting earthquakes is one of the most important 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.”

Description of the challenge on
<https://www.kaggle.com/c/LANL-Earthquake-Prediction/overview>

1.1 The platform

Kaggle is one of the most well known platforms for machine learning challenges and competitions. Third-party members like universities, private companies, and even governments can submit challenges to the public. These consist of a description of the problem, the data and, optionally, a submission timeline.

Kaggle as a platform provides the tools for downloading the data, sharing your approaches/solutions with others and even allow the end-user to implement their solution in an easy to understand way. The community character in particular is special: one can ask a question in the forum, for example, and get an answer from other members of the community quite quickly. Sometimes, these answers are also of high quality and created by professional data-scientists.

1.2 The data

The data can be downloaded from Kaggle after registering and accepting the Honor code of the project.

The data consists of approximately 8.9 gigabytes of Comma-Separated-Value (CSV) files with realistical, acoustical measurements before earthquakes.

The training data contains continuous measurements of acoustical data. , hist

acoustic_data,time_to_failure	acoustic_data
12,1.4690	6
6,1.46909	5
8,1.46909	3

Figure 1.1: train.csv

seg_id,time_to_failure
seg_00030f,0
seg_0012b5,0
seg_00184e,0

Figure 1.2: test.csv

Figure 1.3: example_submission.csv

As we can see in Figure 1.1, the features are one-dimensional and discrete numbers (*acoustic_data*), while the target is a continuous number (*time_to_failure*). So, this challenge is basically a regression problem with one input and output.

The size of the data also increases the difficulty as training/test/validation phases get longer. Additional measures must be taken for account for these longer waiting times, eg. automating

Field name	Type	Description	Example value
acoustic_data	Input	“the seismic signal [int16]”	12
time_to_failure	Target	“the time (in seconds) until the next laboratory earthquake [float64]”	1.4690
seg_id	Submission input ID	“the test segment ids for which predictions should be made (one prediction per segment)”	seg_00030f

Figure 1.4: Descriptions taken from the challenge website

the testing of different algorithms in parallel instead of waiting for the individual evaluations to finish sequentially and starting a new one.

An additional difficulty arises due to the fact that the features are one-dimensional. So, one has to be careful which algorithms to choose instead of just letting a very big neural figure it out semi-automatically and just waiting for the grid/random search to be over.

1.3 The opportunity

The especially prevalent character of sharing in the Kaggle community results in a beautiful learning opportunity. Since we do not have enough time beside our studies, this course would give us a great way to engage in “real” problem solving - with real competition and collaboration of an eager machine learning community and a trove of knowledge in form of a highly visited forum of the challenge. Together with the enforced best-practices of the Kaggle platform, eg. never obtaining the solutions for the test set - only calculated scores, this challenge may also *challenge* us not only to learn more machine learning but about scientific practice itself.