Predicting Earthquakes

Exposé for the "Deep Learning: Architectures and Methods" project

David Gengenbach (david_marlon.gengenbach@stud.tu-darmstadt.de ↗)

Tom König (tom.koenig@stud.tu-darmstadt.de ↗)

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Fachbereich Informatik

Abstract

For our "Deep Learning: Architectures and Methods" project we plan to participate in an online machine learning challenge called **LANL-Earthquake-Prediction**⬀ hosted on the Kaggle⬀ platform. Here, we try to predict when a next earthquake will happen 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."

<div align="right">

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

</div>

The objective is to use historical, acoustic data to predict **when** the next earthquake will happen.

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 (real) historical, measurements before earthquakes.

The training data contains continuous measurements of acoustical data. The test data consists of several approximately 2600 different continuous measurement sets without the target value (*time_to_failure*).

As we can see in Table 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 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.

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 1.1: Descriptions taken from the challenge website

1.3 The approach

Since this is a (linear) regression problem and we are not that knowledgeable of algorithms in this domain, the challenge will call for a lot of learning on our side. Our hope is that we can get to a good approach by looking at previous submissions and forum posts on Kaggle. This will, hopefully, result in a novel approach.

1.4 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 about machine learning but about scientific practice itself.