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.

It contains approximately 8.9 Gigabyte of Comma-Separated-Value (CSV) files with real, historical data.

Examples are shown below.

```
acoustic_data,time_to_failure
12,1.4690999832
6,1.4690999821
8,1.469099981
5,1.4690999799
8,1.4690999788
8,1.4690999777
9,1.4690999766
7,1.4690999755
-5,1.4690999744
```

Listing 1.1: Example values from the train set, *train.csv*

```
acoustic_data
6
5
3
5
6
3
2
3
7
```

Listing 1.2: Example values from the test set, *seg_ff7478.csv*

```
seg_id,time_to_failure
seg_00030f,0
seg_0012b5,0
seg_00184e,0
seg_003339,0
seg_0042cc,0
seg_004314,0
seg_004cd2,0
seg_004ee5,0
seg_004f1f,0
```

Listing 1.3: Example submission file, *sample_submission.csv*

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.