Machine Learning Engineer Nanodegree

Capstone Proposal

Kyle McMillan

January 13th, 2018

Earthquake Prediction

Domain Background

In September 2010 at 4:35am local time, there was a large magnitude of 7.1 earthquake in Christchurch, New Zealand. This caused widespread damage throughout the city but there were luckily no fatalities linked to it. On February 22nd at 12:51pm there was another large earthquake, this time measuring 6.3. This cause even more destruction to an already devastated city and cause the deaths of 181 people. While this is just one case of one city, there are many cities all over the world that have been hit by large earthquakes. Larger ones don't only affect one city and can leave problems for the whole country, or even other countries.
Every year, earthquakes case millions of dollars in damages to property, but more importantly, earthquakes can have high death tolls. For example in 2004 there was 298,101 deaths worldwide from earthquakes (https://www.statista.com/statistics/263108/global-death-toll-due-to-earthquakes-since-2000/).
With the ability to be able to predict when an earthquake may appear can give people time to move. If there is a tsunami involved, such as in the Indian Ocean earthquake (https://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake_and_tsunami), an early warning can give people time to move to higher ground. Or give an incentive to build stronger structures in an area where a strong earthquake could be predicted to appear.
For this project, a single country's earthquake information will be used. New Zealand is one country that is located in the Pacific Ring of Fire and has thousands of earthquakes per year. In recent years, they there have been a few large earthquakes that, if predicted, could have prevent the death of many people.

Problem Statement

In general, it is believed that earthquakes are impossible to predict, but as computers and machine learning algorithms become more and more powerful it maybe possible to see patterns in the data. By using one country's earthquake data, the overall aim of this project is to see if there is a correlation between the timing, depth, and location (longitude and latitude) that lead to a large earthquake (magnitude 5 or above).

Datasets and Inputs

The dataset will be downloaded from GeoNet (http://www.geonet.org.nz/). All the downloads are free and this project will use the All Quakes in CSV format download.

The raw dataset contains 587,333 (as of 13/01/2018) earthquakes with 21 features. The data will be cleaned up so that all events with in the "eventtype" are listed as earthquakes.

Raw data
Data points with magnitude less than 5: 583,749
Data points with magnitude greater than or equal to 5: 3,584
Event rate: 0.0061 or 0.6%
Since the data is imbalanced, it will require special techniques, such as over-sampling, to compensate for the low event rate.

The features that are planned to be used for predictions in this project are:

- origintime – Time of when the earthquake took place
- latitude – Latitude of the earthquake
- longitude – Longitude of the earthquake
- magnitude – Strength of the earthquake – This will converted to represent values above and below a magnitude of 5 as 0(<5) and 1(≥5).
- depth – How far underground the earthquake was located (in km).

Features to be removed are:

- publicid – This is the unique reference ID of earthquake.
- eventtype – For this project, all of these will be "earthquake".
- modificationtime – If the information to the data point was updated, this time was recorded.
- evaluationmethod, evaluationstatus, evaluationmode, originerror, depthtype and earthmodel – These refer to how the earthquake was evaluated and what computer techniques were used.
- usedphasecount, usedstationcount, minimumdistance and azimuthalgap – These refer to how the location of the earthquake is calculated by the stations.
- magnitudetype, magnitudeuncertainty and magnitudestationcount – These refer to the calculation of the magnitude by the stations.

The dataset manual can be downloaded from https://www.geonet.org.nz/data/types/eq_catalogue

Solution Statement

This project will utilize machine learning techniques to determine if it is possible to predict large earthquakes. The data set being used currently contains over 500,000 data points and in order to fully use this the project will be trained using a random forest. A random forest model is effective at working with large data and also splits the data in training and testing sets.

Benchmark Model

A simple Naive Bayes algorithm will be used as the benchmark. The reason for this is because it is also a classifier type the same as a random forest algorithm. A Naive Bayes algorithm is also a good all-round classifier.

Evaluation Metrics

Since the output of this project will be to investigate the prediction of earthquakes with either a magnitude of greater or less than 5, and the data is also imbalanced, an f-score evaluation will be used.

Project Design

The first stage of this project is to clean up the raw data. Since the raw data contains some points that are not related to earthquakes, avalanches near monitoring stations for example, need to be removed. There are also some points that have no monitoring points, but rather felt by people, these too will be removed. There are also a lot of features that will not be used for this project.
The second stage is to setup and run the machine learning model. Because this will be using a large dataset and the computer used to conduct the learning is not very powerful, this stage will be started early in order to make sure that there is enough time for the model to run.

In the third part, the results of the model will be evaluated and analysed. The documentation and conclusion to the project will be drafted before final submission.

Programing language:
Python 2.7

Python packages:
Scikit-Learn - Open source machine learning library.
Pandas - Open source data analysis.
Numpy - Scientific computing library for python.
Matplotlib Basemap - Library for plotting map points in python 2.7.