Classification Project

Earthquake Data Analysis

Deadline : 6th September 2020, 11:55 PM

1 About

You are given the enclosed data files in csv. They are data about occurrences of earth quakes in a geographical region. The meta data is :

- Sl. No.: Serial Number
- Year, Month, Day: Date of a particular earthquake as per UTC (Coordinated Universal Time)
- Origin Time of earthquake in UTC and IST (Indian Standard Time) in [Hour: Minute: seconds] format
- Magnitude of Earthquake: There are a different way to represent the magnitude of an earthquake. For your study, you can consider Mw, since we are deriving other types from Mw only.
- GPS Location in terms of Latitude(Lat) and Longitude(Long) of earthquake
- Depth: Depth of occurrence of an earthquake in kilometre
- Location: Name of a region where an earthquake took place
- Source: The agency from which we have gathered the data, for e.g. IMD= Indian Meteorological Department, Min. of Earth Science, Government of India

A sample row :

52165	2016	7	7	22:24:02		3.3	3.3	3.164855	2.438576	3.019937

26.8 N	89.5 E	40	Jalpaiguri,West Bengal	IMD

read as:

“A 3.3 magnitude earthquake occurred on 7th July 2016 at 22:24:02 (UTC). The location of earthquake event was Jalpaiguri, West Bengal area with GPS location 26.8 N 89.5 E at a depth of 40km published by the source IMD”

2 Tasks

Use **KNN** and **Decision Trees** to build a classifier for predicting labels as Magnitude(Mw). Use an appropriate threshold that seem fit between [4, 5] inclusive. Consider a threshold of T. For Mw $< T$, label becomes 0 (no earthquake) and for Mw $\geq T$ becomes 1 (earthquake). Use appropriate features as input from the dataset that you seem fit. Please mention what threshold you used and what is the train-test split size that you used in the submission.

1. Plot ROC for both these classifiers for **K** as parameter in **KNN** and **pre-prune depth** as a parameter in **Decision Tree**.
2. Which is the better classifier for this data amongst the two? Give Reasoning.
3. What could be the best possible values of the parameters for respective classifier based on the ROC curves? Give Reasoning.
4. If you have to choose only a subset of two features to predict earthquake, which ones would it be? Give Reasoning. [Hint: You may use nodes of estimated Decision Tree or other techniques]
5. Consider test results of the best model from above analysis. Report the input features that was used to achieve this. Try to improvise the test results by applying feature processing(You may come up with additional features by processing original ones). Report the new set of features that was used and also report the improvements in test results that was achieved. Please use appropriate metrics to report the results.

Submit a pdf file of the results and analysis.

Note : You might be required to remove some columns or rows, or fill in with some default values, in case of null entries in the table. It is left to your discretion. Please feel free to use any tools/libraries that you want to use.

3 Suggestions

The aim of the this project is not just for understanding and implementing classification models , but to analyse how do each of them work and make significant observations based on your analysis . You are free to head in whatever way you desire to present your analysis but below listed are certain recommendations .

- Explain the problem you are trying to solve, short e.g., Using attributes x,y,z, we try to predict Mw.
- Try to explain in your report how did you perform cleaning on the data .

- Required analysis for the above tasks respectively. You can try evaluating the classifier maybe in terms of accuracy, error, recall, etc .
- A few observations that you have made on basis of the analysis .
- Make sure the graphs and labels on the graphs are properly visible.

4 Submission

- Naming convention : <RollNumber1>_<RollNumber2>_classification.pdf, e.g., 20171117_20171185_classification.pdf
- Submission from one member is sufficient .
- This is an open ended project, submissions are hence expected to differ quite a bit . So please refrain from any kind of plagiarism, else cases involved will be dealt with strict penalty .