https://www.kaggle.com/c/predict-volcanic-eruptions-ingv-oe/discussion/190766

What is the sampling rate of the raw and what is the time units?

Both are based on 100th of a second, so the times to eruption in the training data set vary from one minute to 5 1/2 days.

**I see that it says that the observations are 10 min long, there are 60,000 rows per segment. 600 seconds per second. Meaning that the observations must be every 1/100 of a second.**

**Probability distribution (5 pts)**

* Use probability distributions to describe some selected variables.
* What are the parameters in the distribution? Estimate these parameters.
* Use hypothesis testing to show some interesting conclusions.

[**https://contactsunny.medium.com/linear-regression-in-python-using-scikit-learn-f0f7b125a204**](https://contactsunny.medium.com/linear-regression-in-python-using-scikit-learn-f0f7b125a204)

**A**fter that you have to make sure all your features are in the same range for the model so that one feature is not dominating the whole output; and for this, you need [feature scaling](http://blog.contactsunny.com/data-science/why-do-we-need-feature-scaling-in-machine-learning-and-how-to-do-it-using-scikit-learn). Finally, [split your data into training and testing sets](http://blog.contactsunny.com/data-science/how-to-split-your-dataset-to-train-and-test-datasets).

Because our data is all numbers and there’s no text in it, we don’t have to label encode or one hot encode our data. And because we’re using the LinearRegression class, we don’t even have to worry about feature scaling, as this is taken care of by the library itself.

**Split data into Test, train, Validation:**

<https://datascience.stackexchange.com/questions/15135/train-test-validation-set-splitting-in-sklearn>

Could just use sklearn.model\_selection.train\_test\_split twice. First to split to train, test and then split train again into validation and train. Something like this:

X\_train, X\_test, y\_train, y\_test

= train\_test\_split(X, y, test\_size=0.2, random\_state=1)

X\_train, X\_val, y\_train, y\_val

= train\_test\_split(X\_train, y\_train, test\_size=0.25, random\_state=1) # 0.25 x 0.8 = 0.2

**What is meant by guassian curve:**

Normal distribution, also known as the Gaussian distribution, is a [probability distribution](https://www.investopedia.com/terms/p/probabilitydistribution.asp) that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean. The standard normal distribution has two parameters: the mean and the [standard deviation](https://www.investopedia.com/terms/s/standarddeviation.asp). The shape of the normal distribution is symmetric and unimodal.

The model fits theoretical distributions to observed duration data and relies on past eruptions being a good indicator of future activity.

The null hypothesis of this test is that the observed sample can be said to have derived from the theoretical distribution being tested.

**Null hypothesis:**

Two sensors having high correlation results in the same time of eruption.

mu1 = mu2 ?

Two sensors having low correlation result in different time of eruption

mu1 not equal to mu2 ?

https://www.youtube.com/watch?v=Nz4WB8-gNBg&t=475s

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

Which models: Why and which is better?​Given the answers to question 1 and 2, what evaluation metric would you choose to validate that your algorithm is "good enough"?

SVC model:

* Effective in high dimensional spaces.
* capable of performing binary and multi-class classification on a dataset