Machine Learning Final project

DengAI: Predicting Disease Spread

TAs
ntu.mlta@gmail.com

Description

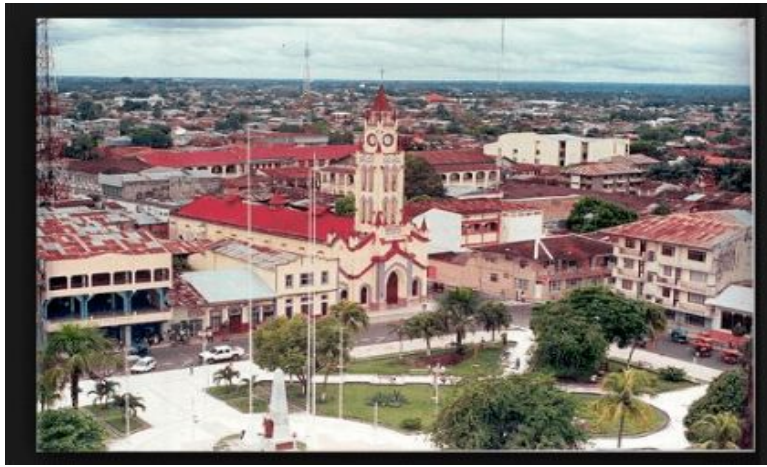
Dengue fever is a mosquito-borne disease.
Because it is carried by mosquitoes, the transmission dynamics of dengue are related to climate variables such as temperature and precipitation.

Competition Website

Environmental data collected by various U.S. Federal Government agencies—from the Centers for Disease Control and Prevention to the National Oceanic and Atmospheric Administration in the U.S. Department of Commerce

Goal

Your goal is to predict the total_cases label for each (city, year, weekofyear) in the test set.
There are two cities, **San Juan and Iquitos**, with test data for each city spanning 5 and 3 years respectively.



Data - External data is not allowed

Around 1500 training data and 400 Testing data
(San Juan train 1990~2008 , test 2008~2013)
(Iquitos train 2000~2010 , test 2010~2013)

Four files :
dengue_features_train and **dengue_labels_train** are indexed by (city, year, weekofyear).

dengue_features_test and **submission_format** are indexed by (city, year, weekofyear).

Data - dengue_features_train

Each (city, year, weekofyear) contains 20 attributes of climate data.
Temperature, precipitation, humidity, vegetation index

Evaluation

Performance is evaluated according to the mean absolute error.

Official Baseline - Include Code

hypothesize that the spread of dengue may follow different patterns between the two cities, divide the dataset, train seperate models for each city

- Handle Missing data
- Analyze correlation between features and label
- Useful Statistics and visualization of data

Negative Binomial Regression

Data - dengue_labels_train

label for training features

Data - dengue_features_test

Same format as dengue_features_train

Data - submission_format

The example of your submission to DrivenData.
Your job is to fill the total_cases colmn.u
Keep in mind that you need to submit one csv with predictions for both cities!