Capstone Project Proposal

Introduction

This document outlines my research proposal for the Machine Learning Nanodegree Capstone Project. It presents my choice of machine learning problem and data sources together with my motivation for this choice; my anticipated approach to the problem; the way I intend to evaluate my models and the amount of effort to be invested.

Domain background

As a resident of the United Kingdom it is easy to forget about the existence of serious infectious diseases which have been largely eliminated here. However they are still a cause of great hardship and economic drag in many places and with changing climate patterns over the coming decades these diseases are likely to spread. One such disease is Dengue Fever which is common in Latin America and South East Asia, infecting up to 400 million people each year. Dengue fever is mosquito borne and therefore liable to spread as climate change increases the habitable range of these insects. Information about dengue fever and its epidemiology can by found on the Centre for Disease Control website at:

https://www.cdc.gov/dengue/index.html

Early warning systems for dengue outbreaks based on monitoring internet search terms have been tried in several places including China, see:

http://journals.plos.org/plosntds/article?id=10.1371%2Fjournal.pntd.0005354

Statistical / machine learning approaches have also been developed, over both short and medium timescales, for example

https://ij-healthgeographics.biomedcentral.com/articles/10.1186/1476-072X-14-9

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0152688

Dengue fever is a widespread and serious health issue which it would be very satisfying to make a small contribution to alleviating, or at least gain some better insight into.

Problem Statement

Infectious diseases like dengue fever are by nature episodic and it is usually not cost effective to be constantly on standby in case of an outbreak. The problem is to predict as accurately as possible the number of dengue fever cases expected, based on currently collected environmental data, so that resources can be more efficiently and effectively allocated. This is a multivariate regression problem.

Datasets and Inputs

This problem is being run as a competition by the “Driven Data” website, see

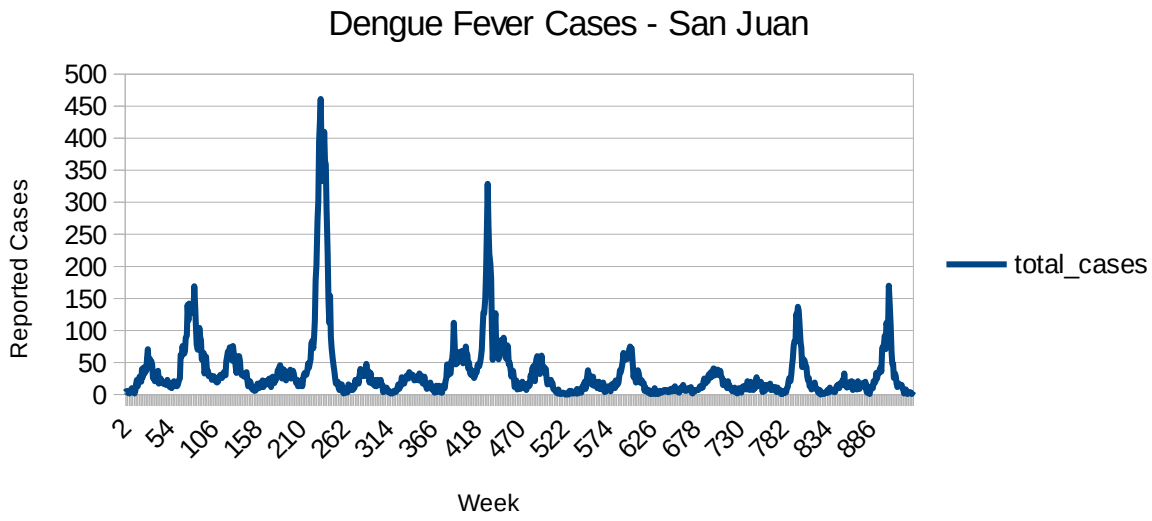
https://www.drivendata.org/competitions/44/dengai-predicting-disease-spread/

The competition provides datasets with weekly environmental measurements, such as temperature, humidity etc, over a 10 year period for 2 cities, together with the number of reported dengue fever

cases for each week. As a mosquito borne disease, dengue fever trends in a given location can be expected to follow mosquito numbers, which in turn will depend on local climate conditions, albeit in a complex way. See for example this article

http://ehp.niehs.nih.gov/wp-content/uploads/121/11-12/ehp.1306556.pdf

Looking at the label dataset for San Juan, containing the number of reported dengue fever cases each week, there is a clear seasonal trend in the data. There are also frequent outbreaks but these do not occur every season, so we need to predict more than just the seasonal variations.



The features dataset includes weekly measurements of 20 environmental features, for example:

station_max_temp_c	33.3
station_avg_temp_c	27.7571428571
station_precip_mm	10.5
station_min_temp_c	22.8
station_diur_temp_rng_c	7.7
precipitation_amt_mm	68.0
reanalysis_sat_precip_amt_mm	68.0
reanalysis_dew_point_temp_k	295.235714286
reanalysis_air_temp_k	298.927142857
reanalysis_relative_humidity_percent	80.3528571429
reanalysis_specific_humidity_g_per_kg	16.6214285714
reanalysis_precip_amt_kg_per_m2	14.1
reanalysis_max_air_temp_k	301.1
reanalysis_min_air_temp_k	297.0
reanalysis_avg_temp_k	299.092857143
reanalysis_tdtr_k	2.67142857143
ndvi_location_1	0.1644143

ndvi_location_2	0.0652
ndvi_location_3	0.1321429
ndvi_location_4	0.08175

Solution Statement

Prediction of the number of dengue fever cases will be attempted by fitting a non-linear regression model to the environmental data and reported number of dengue fever cases from the two cities in the data set provided in the competition. The model will be fitted and tuned separately for each city using a combination of original and composite features, and tested against a hold out set. Initially I intend to use a random forest regressor with both current and past feature data.

Benchmark Model

The competition provides a benchmark model which can be found at the following link;

http://blog.drivendata.org/2016/12/23/dengue-benchmark/

This benchmark model first fills in missing data by simply duplicating prior measurements and then selects the 4 features that are most strongly correlated with the observed number of dengue fever cases. The data is split into a separate training and validation set for each city. A negative binomial regressor is then fit to each training set for values of the ancillary parameter alpha varying from 10e-8 to 10e-3. The negative binomial distribution describes the number of successes in a series of trials that terminates after a predetermined number of failures. The alpha ancillary parameter is the inverse of the number of failures before the trial series terminates. The resulting models are used to make predictions on the validation sets and the best value of alpha is selected based on the mean absolute error of the predictions. The mean absolute error of these models is 22.08 for San Juan and 6.47 for Iquitos.

Evaluation Metrics

The evaluation metric for this problem is set by the competition rules to be the Mean Absolute Error. This is a very simple metric which is easy to understand. However I intend to also calculate some other metrics that are less forgiving of large prediction errors. The main use of these predictions would be allocation/re-allocation of resources when dengue fever outbreaks are expected. Most of the time the predicted and actual number of cases will both be small enough that no action is taken and small errors in these predictions will have no impact on decisions. Large errors, where an outbreak is not predicted, or a predicted outbreak does not occur, will have a negative impact on decisions; so reducing these large errors should be the priority. I think that the mean squared error would be a better metric to optimise, or perhaps the R2 score.

Project Design

The methodology for this project will take into account both the nature of the available data and some (limited) domain knowledge. The feature data includes 20 features, but based on the feature names it seems likely that there will be strong correlations between some of them, such as dew point and relative humidity. There are also several variables whose name begins with “reanalysis”, so these are likely correlated with the raw measurements if they are also included, such as “station_avg_temp_c” and “reanalysis_avg_temp_k”. Where two features appear to be measuring essentially the same underlying

variable we can probably eliminate one of them. The data has quite a few gaps which will need to be filled. Since the measurements are a time series taken at regular intervals these gaps could be filled by simple interpolation. However, if there is a strongly correlated variable where a measurement is present it may be better to use this information. To determine this I could also look at how feature values correlate with the value that would be interpolated from previous/following measurements. Once I have filled in missing values I will remove features that seem to be redundant, based on correlation with other features and domain knowledge. Removing some features at this point will enable me to create new features without the dimensionality getting out of hand.

Dengue cases will be strongly affected by mosquito populations. The mosquito has a life-cycle of about 2 weeks and populations are likely to develop over periods longer that 2 weeks due to feedback, so we will likely need to include a significant amount of past data and use a non-linear model to be able to make good predictions. I think that the selection of suitable features, including how far into the past to go and whether to include feature combinations, will be a major part of the challenge. To this end I will choose one or two regression models and examine their performance using a range of features including measurements from 1 to 8 weeks in the past, moving averages of measurements over periods of up to 8 weeks and perhaps polynomial feature combinations.

I plan to start with a Random Forest Regression model. I have chosen this for several reasons:

1) Decision trees are robust to input feature scaling and to non linear relationships between features and labels. Dengue fever outbreaks are clearly a non linear phenomenon

2) Decision trees are good at ignoring features that are not important. If I create a lot of combined features there are likely to be a lot that are not useful.

Depending on how well this performs I may also try a support vector regression. Support vector regression can perform well on non linear relationships using the kernel trick and are tolerant of small errors which suites out goal well.
A simplified project workflow is visualised below.