Dean Gladish
Jordan Aron
Will Thompson

December 14, 2017

Nicholas Reich
Associate Professor at The Reich Lab
University of Massachusetts-Amherst

The aim of this document is to provide a starting point for future externs. Included are definitions and concepts, some specific to the Reich lab and some not, that will aid in understanding the goal of the lab.

## Definitions

- Cross-Validation - Model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set
    - For k-fold cross validation the data is divided into k sections (in this example each section would be one season of flu). The model is then trained on k-1 sections and tested on the one section that was not included in the training set. This process is repeated k times, where each section is given a chance at being the tested set. At the end all of the model assessments are averaged.
- Weighted Influenza-Like Illness (wILI) - Average proportion of doctor visits with influenza-like illness for each state in the region, weighted by state population
    - Useful metric for determining onset and incidence
- Ensemble Model - combination of multiple models to obtain a single prediction that leverages the strengths of each model
    - Each individual model is given a weight that determines its importance in the ensemble model. This weight is determined by the individual model's log score (i.e. a higher log score indicates a more accurate model, which in turn translates to a higher weight).
    - Weights sum to 1
    - Good choice for noisy, complex, and interdependent systems that evolve over time. For instance in the beginning of a season simpler models based on season histories are more accurate, while later in the season models that take into account the current season's facts become more useful
- Stacking - An approach to linearly combine different component models to create an ensemble model.
    - Each component model is first trained and cross-validation to obtain measures of performance
    - A stacking model is trained using the cross-validated performance measures to learn how to optimally combine the models
- Log Score - a scoring rule that measures the likelihood that a discrete bin contains the true observation. In context, it is the natural log of the probability estimate for the actual outcome.
- Kernel Conditional Density Estimate - an estimated composite distribution that is the result of stacking.
    - The bandwidths of the Gaussian distributions that are being stacked are determined by the closeness of their targets to the target at hand
    - An estimation based on a bunch of Gaussian kernels.
    - Similar to KDE however the current season's data is used (hence conditional on the history of the future season). This differs from KDE as KDE makes one prediction in the very beginning of the season and does not change it.

- Degenerate Expectation-Maximization Algorithm - an algorithm that, in this case, determines weights to allocate to different probability models (based on their likelihood of being correct) for the purpose of creating an ensemble model.
    - This algorithm uses the log scores, giving models with a higher log score more weight in the ensemble
- Gradient Tree Boosting - a machine learning technique that creates a function designed to minimize a given loss function.
    - A machine learning method that uses a forward stagewise additive modeling algorithm to iteratively and incrementally construct a series of regression trees, the sum of which minimizes a given loss function.
- Incidence - The wILI at some point in the season
- Onset Week - the first of three successive weeks of the season for which wILI is greater than or equal to a threshold that is specific to the region and the season

## Prediction of infection disease epidemics via weighted density ensembles

## Ensemble Model

______This paper discusses the use of ensemble models in modeling influenza in the United States. Ensemble models combine results from multiple component models, leveraging the strengths of each one, to create a more accurate prediction. Three approaches were taken to create the ensemble model: i) equally weighted average of all models, ii) constant but not necessarily equal equal weights for all the models, and iii) model weights that change over time (i.e. the model weights depend on what point in the season it is).

## Component Models

Kernel Density Estimation (KDE) - This is the simplest of the component models. KDE is a non-parametric (i.e. no easy description by a probability distribution) estimation method. This model makes one prediction at the beginning of the season based on the past history, and does not updated predictions as more information is made available, therefore this model is more accurate in the beginning of the season when less information is available and we must rely on past seasons.

Kernel Conditional Density Estimation (KCDE) - This method gives conditional distribution for incidence at one future time point given recent observations of incidence and the current week of the season. It acts very similarly to KDE except that it uses this season's information to update predictions.

Seasonal autoregressive integrated moving average (SARIMA) - Includes seasonal (S), autoregressive (AR), integrated (I), and moving average (MA) concepts. The seasonal term accounts for the seasonality aspect of influenza and allows the ARIMA portion to function properly. The autoregressive aspect attempts to deal with the dependence of observations close to each other (as this is a time series two points close in time will have close values). The integrated aspect deals with the non-stationarity, and the moving average portion allows the modeling of a parameter that changes (i.e. incidence based on week).

It additionally includes a first-order seasonal differencing operation in order to account for the seasonality aspect of flu-like illness and allow the ARMA portion (which assumes that the time-series is stationary) to function properly. An ARIMA model is a generalization of the auto-regressive moving average model, which involves regressing a variable based on its past values (hence the auto-regressive portion) and the implementation of a moving average of various subsets of data in order to predict the future. The auto.arima function in the forecast package in R chooses the differencing, auto-regressive, and moving average terms.

## Methodologies for Ensemble Weights

1. Equal weights
2. Constant model weights (but not necessarily equal) weights by degenerate EM
3. Feature-weighted (FW)
    a. Weights change based on features such as time of the season and uncertainty of the model
4. Feature-weighted with regularization
    a. Similar to FW but discourages extreme values and large fluctuations

## Misc.

Predictions can be broken down into two general areas: seasonal and forecasting. Forecasting predictions predict the incidence one, two, three, or four weeks in the future. Seasonal predictions are: peak incidence, peak week, and onset week.

One important thing to note is that sometimes there is a difference between the reported data and the true data, which is not corrected for some time. This leads to predictions based on the reported data, rather than the true data. In turn, this can lead to gross errors in prediction.