Dean Gladish
Project Outline and Description
June 13, 2018

Quantifying similarities between models used in the Ensemble

As we know from Professor Nick on Monday, we don't yet have a clear *single* metric for quantifying model similarities. He suggested that we could take pairs of predictive distributions from the ensemble model and use something like the Kolmogorov-Smirnov test to measure their differences.

Given an i.i.d. sample of data from an unknown distribution $\mathbf{P}$ and a proposed distribution $\mathbf{P}_0$, the Kolmogorov-Smirnov test will tell us which of the following hypotheses,

$$H_0 : \mathbf{P} = \mathbf{P}_0,\ H_a : \mathbf{P} \neq \mathbf{P}_0$$

are true. Unlike the χ^2 goodness of fit test, the K-S test is stronger because it doesn't compare discrete groups to their expected contents; it considers each data point on a continuous distribution[1]. Furthermore, the K-S test is easier to implement – the K-S test statistic

$$D_x = sup_n[\,|\,F_n(x)\ -\ F_0(x)\,|\,]$$

is the supremum (over the domain) of the vertical difference between a data point and the proposed distribution[2]. Larger sample sizes allow us to be more certain of its significance.

Since my project is about quantifying similarities between pairs of models, I will use the K-S test to compare each pair of distributions in the ensemble to derive a test statistic. The magnitude of this test statistic will give the extent of their dissimilarity and show which models are truly independent. A further step that will be taken is to use similar tests like Kuiper's test as a way of measuring which models best fit the data, which is another good thing to consider when combining them for the ensemble.

The component predictive distributions for the ensemble were based on KDE, KCDE, and a SARIMA implementation, and their accuracy was measured using the log-score (the natural log of the probability that the model assigned to the correct bin)[3]. My project will focus on the relationships between these models using the K-S test. Additionally, I will look at other

---

[1] https://ocw.mit.edu/courses/mathematics/18-443-statistics-for-applications-fall-2006/lecture-notes/lecture14.pdf

[2] https://newonlinecourses.science.psu.edu/stat414/node/322/

[3] http://reichlab.io/pdfs/publications/ray-density-ensembles.pdf

measures of these relationships due to the fact that the K-S test does not give a sense of the total variation between models over time and may be influenced by outliers. Among these measures will be the sums of squared residuals between the distributions, the AIC criterion, and whether there is an association between the slopes of the distributions. For example, one of the questions that I will look at is whether two models tend to increase/decrease at similar rates over time. The Kolmogorov-Smirnov test and similar tests will be used to quantify similarities between the slopes and changes in slopes of the models.

The following is a rough timeline of the next two weeks to come:

June 11 – June 25:

This time period will be full of preliminary diagnostics and figuring out how we should proceed with finding a quantitative way of comparing trends among models.

1. Read the literature on infectious disease forecasting – cross-validation, wILI as a measure of influenza incidence, stacking, how the "degenerate" EM algorithm looks at log scores, gradient tree boosting, etc.
2. Figure out which packages in R (or otherwise), data about the models from the Reich Lab GitHub page (https://github.com/reichlab), influenza data from the CDC (as well as additional testing data obtained possibly by bootstrapping or other means), and other software will be needed in order to run something resembling a Kolmogorov-Smirnov test between the models.
    a. Find the optimal method for generating test data to see how models behave in a variety of different situations.
3. Create diagnostic plots. Plot the differences between the predicted values of models, calculate their coefficients of determination, and visually assess whether they appear to be associated. Try to perform multiple regression of the SARIMA model on the others, and then conduct a likelihood ratio test to assess which of the two other models follows the SARIMA model more closely. Make residual plots between the KDE and KCDE models and plot the differences in their slopes across time as well and see whether the residuals follow a known distribution. These plots should be included in the final presentation and write-up about the subject.
    a. Although these models were generated via different methods, we want to look at residuals between them and see if these residuals follow a certain distribution whether it is normal or exponential (I say exponential because the models appear to deviate more from each other as time goes on[4]).

[4] http://reichlab.io/