Statistical Learning - Milan, Fall2019

Homework problem 0

This exercise relies on the 2009 Nature paper[1] describing Google Flu Trends. Read the Nature paper carefully, making sure you pay attention to the statistical problem formulations and critically evaluate how they are formulated and solved.

Your goal is to identify two clear flaws in the statistical methodology, and explain how each of these aspects could have been better addressed. Make sure your answers are specific, accurate and concise (and as mathematical as possible), rather than vague statements. The goal here is not to propose completely different approaches and ideas, but to concentrate on how they could have implemented their chosen approach better.

Here are concrete hints for flaws you may consider (though you can find others):

1. The internal four-fold cross validation in fitting each one of the $5 \cdot 10^6 \cdot 9$ models described at the beginning of the Methods section: is it necessary? Could the same goal have been accomplished with a simpler approach? You may want to go back to your linear regression course notes and think about the meaning of calculating a correlation on the holdout data...

2. The methodology for combining the 36 Z scores into one score described in Methods: is there an obvious better / more powerful approach? Possibly combined with the solution of the previous item.

3. Given that they want to predict the actual percentage in the future, would the use of other evaluation measures in the final step (selecting the number of terms, as described in Figure 1) be more appropriate? Which and why?

[1] http://www.nature.com/nature/journal/v457/n7232/full/nature07634.html