Midterm Exam 1

Advanced Methods for Data Analysis (36-402/36-608)

Due Thursday March 6, 2014 at 11:59pm

Instructions: you will submit this take-home midterm exam in three parts.

1. *Writeup.* This will be a complete writeup, in full data analysis report format, of what you have done. More details on the format to follow. And, as with your homework assignments, your submitted writeup must be in PDF format. You can produce this PDF using LaTeX, Word, or whatever you want, but at the end of the day, it must be a PDF. Any other type of file will not be accepted.

2. *Code.* Also submit your R code for the homework as a plain text file. You should demarcate sections of the code that correspond to different tasks using clearly visible comments (e.g., as in `##### Cross-validation #####`).

3. *Data set.* Each of you will be emailed a slightly different data set, and for completeness you will submit this (Rdata) file as well.

You must not communicate at all with your classmates, your friends, or anybody other than the Professor and the TAs about this midterm assignment, during the week you have to complete it. Evidence of illegal communication will be taken very seriously.

You are of course allowed to consult your class notes, your homework assignments, or any of the other course materials in order to complete this exam. You may use the internet as a resource but you may not use it to communicate with anybody else or ask questions about your assignment. Everything you write must be in your own words, and you must explicitly cite any sources used other than the class materials.

The following describes the scientific problem and statistical questions that you must address, as well as the format of your writeup.

1 Forecasting the flu

The data set that you will be examining for this midterm assignment is based on real data on the flu, from the Center for Disease Control (CDC). Each week, participating doctors across the country report the percentage of patients they see with an influenza-like illness (ILI), in that given week, to the CDC. The CDC weights these percentage ILI measurements, to account for the fact that the composition of reporters changes from week to week, yielding measurements we will call WILI values. In short, you can think of the WILI values as (noisy) indicators of the national incidence of flu, across the weeks.

You should have been emailed an R data file `fluxxx.Rdata` (where `xxx` stands for a number between `001` and `100`). When loaded into your current R session, this gives you a list named `flu`, with three components: `season`, `week`, `wili`. The `wili` component contains the CDC's provided WILI values, for each week, over the last 11 or so years. The components `season` and `week` provide a way of annotating the time aspect of these WILI values. The `season` component has entries between 1 and 11, which denote the season—a 52 week period spanning about week 21 of one calendar year

to week 20 of the next—associated with each WILI value. E.g., the first 52 entries of `season` are 1, which means that the first 52 entries of `wili` were measured over what we consider the 1st flu season, spanning 2003-2004. The next 52 entries of `season` are 2, which means that the next 52 entries of `wili` correspond to the 2nd flu season, across 2002-2003, etc. Finally, the `week` component simply reflects the week underlying each WILI value, and for simplicity, the entries of `week` within a given season have been recorded between 21 and 72 (rather than wrapping back around to 1 after week 52). E.g., the first 52 entries of `week` are

$$21, 21, 22, \ldots 52, 53, 54, \ldots 72,$$

where the values $53, 54, \ldots 72$ actually correspond to weeks $1, 2, \ldots 20$ of the 2002 calendar year, but have been offset for simplicity.

Importantly, note that the last season, season 11, has only 29 observed WILI values, and not 52 like all of the others. This is because season 11 corresponds to the current flu season, 2013-2014, and the data you will be analyzing was collected at week 49 of 2013 (so that only the WILI values across weeks $21, \ldots 49$ of 2013, a total of 29 values, are available to you). Your job in this midterm exam will be to forecast the coming flu season, i.e., to predict the WILI values for weeks 50 through 72 of season 11. You *may not go online to try to search for the WILI values for the remainder of the 2013-2014 season*, not that this would actually help you at all.

(Note: do not be intimidated or confused by the word "forecast"—this does not mean anything special, really. If it makes you more comfortable, you can simply substitute "predict" or "prediction" for this word wherever appropriate.)

2 Statistical questions, data analysis report

As mentioned at the beginning, you submit a data analysis report, as well as your R code. The data analysis report can be a *maximum of 5 pages*, and must abide by the section structure described below. Sections 2, 3, 4 are the statistical "meat" of your report, and address Questions 1, 2, 3, to be discussed. You should clearly lay out what you have done, using figures to supplement your explanation. As with the homework assignments, your figures must be of proper (large) size with labeled, readable axes. You should not mindlessly paste raw R output into your writeup with 12 significant digits, etc. You can include R snippets at your discretion, if you think that will help your explanation. In general, you should take pride in making your report readable and clear. You will be graded both on statistical content and quality of presentation.

Section 1: Introduction

The introduction can be brief (1 or 2 paragraphs) but must properly describe the data set and motivate the problem. You cannot copy motivation text verbatim from this document or anywhere else for that matter. Everything you write must be in your own words (and as mentioned previously, this obviously holds for the entire report). You do not need to perform basic exploratory data analysis like you learned in 401, and in fact if you do so here, you will be wasting space. But, e.g., a plot or two of some flu curves could be helpful in portraying the setup.

Section 2: Forecasting season 11

Question 1: *what is this season's flu curve going to look like?*

For each of seasons 1 through 10, you can plot the curve of WILI values across weeks $21, \ldots 72$. For season 11, we only have a partially observed curve of WILI values, over weeks $21, \ldots 49$. The first question you will consider is: what is this last curve going to look like over weeks $50, \ldots 72$? As a first step, you will fit a nonparametric smoother to the 10 past flu curves. I.e., considering each

of the last 10 seasons as a separate data set, you will fit a smoothing spline estimate to the WILI values (treated as the outcome y), as a function of the underlying weeks (treated as the predictor x). Use the following routine to tune the level of smoothness:

- Choose the degrees of freedom value d of the fit by leave-one-out cross-validation.
- Refit the smoothing spline estimate to have degrees of freedom $0.75d$.

This method may seem ad hoc, but often leave-one-out cross-validation does not provide us with smooth enough fits, and this is a way of increasing the amount of achieved smoothness. (We will formalize a better method in the future, called the "one standard error rule".) Hint: the `smooth.spline` function in R is able to perform leave-one-out cross-validation internally, so you shouldn't have to write your own code for this part. Read the documentation for `smooth.spline` (and note that generalized cross-validation and cross-validation aren't the same thing; you want to perform the latter).

Now you have a smoothed curve for each of seasons 1 through 10 (and to emphasize, each of these smoothed curves were fit and tuned separately). Next you will compare each of these curves, across weeks 21 through 49, to the observed WILI values from season 11. In particular, for each season between 1 and 10, compute the squared error between the season's smoothed flu curve (i.e., the fitted values from the smoothing spline estimate) and the observed WILI values, averaged over the weeks 21 through 49. Mathematically, for a season $s = 1, \ldots 10$, this is

$$e_s = \frac{1}{29} \sum_{w=21}^{49} \left(y_w^{(11)} - \hat{y}_w^{(s)} \right)^2$$

where $y_w^{(11)}$ is the observed WILI value for season 11 at week w, and $\hat{y}_w^{(s)}$ is the fitted WILI value for season s at week w. The best-fitting past season s_0 is the one that minimizes this average squared error, e_s, $s = 1, \ldots 10$. To predict, or forecast, the WILI values for season 11 over weeks $50, \ldots 72$, you will simply take the fitted values $\hat{y}_{50}^{(s_0)}, \ldots \hat{y}_{72}^{(s_0)}$ from the best-fitting past season s_0.

Section 3: Assessing forecast error

Question 2: *what is an estimate of the test error of your forecast?*

Suppose that you hand your forecast for season 11 to the CDC, and the first question they ask is: how accurate is this forecast? To address their question, you must come up with an estimate of the test error

$$e_{\text{test}} = \frac{1}{23} \sum_{w=50}^{72} \left(y_w^{(11)} - \tilde{y}_w^{(11)} \right)^2,$$

where again $y_w^{(11)}$ is the WILI value for season 11 at week (but note now that $y_{50}^{(11)}, \ldots y_{72}^{(11)}$ are unobserved), and here $\tilde{y}_w^{(11)}$ is the forecasted WILI value for season 11 at week w (that you derived for Question 1).

You should estimate e_{test} using cross-validation. Hint: think about leaving out one of the past seasons $1, \ldots 10$, pretending like you don't observe its WILI values for weeks $50, \ldots 72$ (even though the season is fully observed), and using the remaining 9 season to forecast its missing WILI values for weeks $50, \ldots 72$. You can then record the test error for this forecast.

Section 4: Estimating forecast variability

Question 3: *what is the forecasted peak of season 11, and how variable is this forecast?*

Finally, suppose that the CDC is particularly interested in the forecasted peak for the coming season—this is the maximum WILI value across the entire season. (Remember that this reflects the maximal national incidence of flu across the season). You can read this off directly from your forecast for season 11, by just looking at the maximum fitted value $\tilde{y}_s^{(11)}$, $s = 50, \ldots 72$.

Further, suppose that the CDC is not only interested in such a point estimate, but they also want to know how variable it is. I.e., if the forecasted peak for season 11 is $\tilde{y}_{\text{max}}^{(11)}$, the CDC wants to know

$$\tilde{y}_{\text{max}}^{(11)} \pm \Delta,$$

where Δ is an estimate of the standard deviation of $\tilde{y}_{\text{max}}^{(11)}$. You will estimate Δ using the bootstrap. Hint: you will want to use the bootstrap to resample your entire data set, seasons 1 through 10, and season 11 as well. It is the WILI values that you'll want to resample; the weeks here are fixed and cannot be resampled. Within each of seasons 1 through 11, use the residual bootstrap to resample the WILI values (this means resampling errors about the fitted spline curve—so you'll have to fit a smoothing spline to season 11 as well). Then repeat the forecasting procedure from Question 1 on this resampled data set, and record the forecasted peak. Doing this, e.g., 500 times, you will have a collection of forecasted peaks from the bootstrap.

Section 5: Discussion

The discussion can be brief (1 or 2 paragraphs). You should summarize your analysis. You can use this room to reflect on the results that you view as successes, and the results that raise suspicions in your mind. You can also provide ideas for what you might consider in the future, in this data setting.

3 Warning: Question 3 may be hard

A word of warning is that the bootstrap task in Question 3 may be difficult. You should make sure that you have thoroughly figured out Questions 1 and 2 before you approach Question 3. While it is true that for full marks you have to complete all questions, *you will still be able to get a good grade even if you are not able to produce an estimate* Δ *in Question 3.* Therefore, to repeat, make sure you spend adequate time figuring out Questions 1 and 2, and crystallizing your writeup for these two, before you spend a lot of time on Question 3. And, in the absence of actual results (i.e., if you are having programming difficulties), you can still receive partial credit for Question 3 by explaining your attempted technique in words.