Additional Questions and Answers About Module 1

In addition to the questions on the Discussion board, these questions were e-mailed to me and was last updated on Sept. 14. All the students in the e-mail gave permission to share their questions to the class.

1. *The first is about pattern of input data. As you have mentioned, we will have some missing values and non-scalar values. Yet I want to know if we will have some manual manipulation. For example, in the project description, the sever hard drive data was cleared at around 1100. Will this kind of pattern appear in our input, or it will be just like data between 0 to 1000 or 1200 to 1400?*

Ans: There will be some manual manipulation, like the plot in the assignment. Also, it's important to recognize that there is no "universal" algorithm that can handle every possible data situation; we do expect that some algorithms will perform better than others in certain data settings that we will test on.

1. *The next is about how long we want to predict. You mentioned as an example, in the sever data in the description, you want a warning before two weeks. So will the tail of our input be close to the maxcap, or we have to make a long-term prediction? E.g. in the out\_box example, we cannot predict when it will reach maxcap if only given data before 200.*

Ans: This depends on the time unit used in t + the server engineer's preference. You definitely don't want to overshoot and exceed maxcap, but you also don't want to be too early in your prediction. I gave an example of two weeks, but certainly some engineers may require a month or just a few days. In general, there is no explicit, mathematically defined criterion in practice and for the module, giving a good "ball-park" estimate under maxcap will be sufficient.

Also, with regards to your example, as mentioned in my response to question (1), I don't expect that algorithms will be able to accurately handle the out\_box.csv case where you're only given 1-50 time points; that's why when there are human analytics team to handle these kind of erroneous behaviors.

1. *I’m wondering how simple the algorithm should be. Can we include equations like y\_t = \mu + \beta \* t +  e\_t to specify our model?  And do we need to check the model assumptions to show how the model suit, or it would be enough to be fast, scale, robust and accurate? Will some simple time series methods, like ARIMA, be too statistical to be used?*

Ans: This will depend on making the necessary trade-offs and assessing whether the trade-offs are worth it. For example, if you find that the lm(y\_t ~ t) is more than sufficient in your own tests for prediction accuracy, speed/scale, and robustness, then it makes sense to use this model. However, if you find that lm(y\_t ~ t) is fast and robust, but not as accurate, then it will depend on how much inaccuracy you are willing to tolerate at the expense of a simple model.

Relatedly, ARIMA may be a good approach, especially if its achieving excellent predictive performance at the expense of a bit of complexity.

Finally, for model assumptions, you do not necessarily have to explicitly check for model assumptions. However, some discussion of them as well as their limitations under certain data settings would be useful for the end-user to know when he/she uses your algorithm for prediction.

1. *As for the realistic simulated data, are we supposed to simulate our own data that follow the pattern in this server context, or to obtain some existing dataset? Are we supposed to also include the performance of our proposed method on two given datasets (YouTube and Box) in the .pdf file?*

Ans: You can choose to include the two given datasets as part of your performance metrics. You can also find your own real or simulated data to try your method; I strongly encourage you to do the latter to really test your method and to understand the limitations of your method. You do not have to report every simulated data you have created and tried, but trying to test your method against many data generating scenarios will help you better understand your method a bit better.

1. *When you check the R code, will you check it on code other than the give .R file?*

Ans: We will use the testing suite code in the given .R file. The only thing that will change is the data that is used inside of the given .R file. In particular, we will test your method across different datasets, some of which will be similar to the ones on Canvas, while others that may be not so similar.

1. *Should we deal with the missing data for the raw data we have ? In the lecture, professor you taught us a small trick in R that is toupper() == tolower() and it seems that this code will also directly drop those NA data. But I think maybe we can't directly throw the NA data because some of them may not miss by random and we could perhaps fill these data with some interpolations. So, I just want to confirm with you whether we need to carefully deal with those missing data.*

Ans: This is a great question and it depends on your algorithm. But, broadly your algorithm should be robust to missing data (i.e. it should still produce stable estimates even if missing data are present). There are many ways to make an algorithm robust to it; the example that I gave is one particular one based on a hack that I learned some time ago, but there are also many statistical techniques to deal with missing data.