Data Mining Fall 2017, Final Team Challenge

This document contains descriptions for the final team challenge for the Data Mining course. This challenge is to be completed in your chosen teams. The deadline for this assignment is officially November 24, and I will expect your team to make a submission by then. However, you may continue to update your submission as you please until the assignment is graded (sometime around the finals period). The description of the assignment is given below

0.1 Challenge Task

There are two (CSV) files on the moodle, a training dataset `dm-final-train.txt` and a test dataset `dm-final-testdist.txt`. Stats about these files are enumerated below

- `dm-final-train.txt`: 1875 rows, 121 columns, training instances
  - Column 0: index (for bookkeeping, not a feature)
  - Columns 1-100: features
  - Columns 101-120: targets
- `dm-final-testdist.txt`: 625 rows, 101 columns, test instances
  - Column 0: index (for bookkeeping, not a feature)
  - Columns 1-100: features

Your task is to model the relationship between the 100 features and the 20 targets using the techniques and tools we have discussed in this course. This problem involves *time-series.* Both the features and the targets are observations of how a phenomenon changes with time. The goal is to predict the behavior of the phenomenon in the future given how it has behaved in the past. This is illustrated in Fig. 0.1, which shows a test instance along with the true target values. In brief, your model must predict the "blue" samples given the "green" samples.

Your submissions should be formatted in the same way as `dm-final-train.txt` (i.e. you should include both the features and the predicted targets in your file). You are free to use any of the tools we discussed. A portion of the grade will be based on relative performance of the teams. Improperly formatted submissions will get 0.

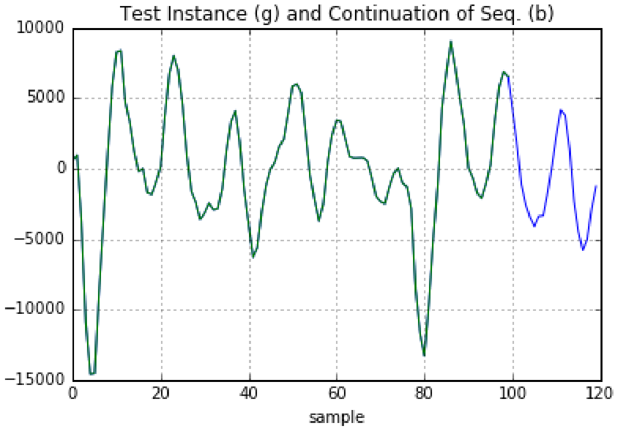

Figure 1: sample test instance (green), and continuation of sequence (blue)