Fundamentals of Data Science
Fall 2017
Daniel Egger

Revised Instructions for Scoring Cisco Test Set Data
November 28, 2017

I have slightly changed the plan for next week. We will provide four (4) cases for the test set. Two (2) cases contain single machine failures, and two have no failures. The data columns and time series format will be the same as the training set files. The background level of network traffic is also the same.

Each case will be provided in the form of a series of .csv files, with the followings structure (times given are approximate)

Time 0 to 15 seconds
Time 0 to 25 seconds
Time 0 to 35 seconds
And so on….

Each case will be given as eight (8) separate files.

After evaluating each file, you will assign reported probabilities to the 13 possible outcomes at each designated time interval. You MAY NOT change the probabilities after examining later files that extend the time series. This would only be possible in a real-time system if you had a time machine!

The whole exercise is to make the best estimate you can with the incomplete information, incrementally improving your estimates as more data become available.

You probabilities will be entered into a standard Excel Workbook – provided separately now.

Remember – since no estimate can be less than 1%, and estimates must always sum to 1, no individual outcome estimate can be greater than 88%. Even if an estimate reaches 88%, you should continue modify that estimate for files containing more data – for example it is possible that you made a mistake.

The first set of test files will be provided on Monday December 4, and your results for that set should be entered into the Excel workbook and uploaded to Sakai Dropbox by Monday 11:45 pm.

The second set of test files will be made available on Tuesday December 5, and results should be entered into a second copy of the Excel sheet and uploaded by 11:45 pm on Tuesday.

You can do scoring on test and training files yourself once you determine the correct categorization for each of of the four cases.

In the row marked "*outcome vector*", enter 0 for all values except the correct outcome. Score the true outcome as 1.

Teams should compare their information gain performance on the test files to performance on the training files *at comparable time intervals relative to the known failures*.

In other words, if a test set file contains a failure and one of the estimates is approximately 20 seconds into the beginning of the failure, then the information gain/loss of that score should be compared to information gain/loss on the training set files containing a failure, approximately 20 seconds into the start of that failure.

A good model should perform approximately as well (within approx. +-25%) on the test set as on the training set, at the same time intervals relative to failures.

The most important goal is to maximize information gain early in the failure – within the first 30 seconds – without over-fitting (false positives) on any of the null data. You must include your information gain performance on training and test files in the written reports due Thursday, December 7. These documents must also contain your exact model algorithm(s), expressed as formulas and if-then statements. You should also discuss any interesting results in your Powerpoint Presentation (remember, 5 minutes maximum, plus time for questions).