Course Project Guideline

This guideline provides you with some basic directions for organizing the data analysis and machine learning tasks of your course project using R with the electricity data provided.

NOTE: A new training data set has been uploaded to the course page. A collection of related test data sets to be used for the course project will be uploaded over the next couple of days.

In the description of your findings, beyond quantitative results, the emphasis will be on your interpretation of qualitative aspects of the findings.

Phase 1. General data exploration

First of all, you need to determine a meaningful granularity for the time window(s) within workdays or weekend days as well as different months or seasons of a year as a basis for your analysis. This aspect of the problem is directly related to the third question of the second group assignment.

APPROACH: Checking for overall changes relative to the expected normal behaviour

(I) Consider checking the test data against the training data with respect to various characteristic features for the identified time windows. A fairly basic check could be done by calculating and comparing basic characteristics such as the mean and standard deviation.

(II) Further, consider various seasonal trends using regression and/or other statistical models.

Phase 2. Anomaly detection approach

APPROACH 1: Finding Point Anomalies.

(I) Out of Range.

Based on the *Min* and *Max* values of a certain feature in the identified time window(s) in the training dataset, one can detect all those values (of the same feature) in the test dataset which are either above or are under the *Max-Min* range (i.e., point anomalies or outliers).

(II) Moving Average.

Step 1: Consider a fixed size window of observations (e.g., a window of 7 observations).

Step 2: For a specific feature, calculate the average of the window and then slide the window by one observation. (This will eventually smoothen the curve of that feature.)

Step 3: At any point of Step 2, if the difference of the value of the observation and the calculated average is either above or below a certain threshold, that observation can be considered a point anomaly of the feature in question.

APPROACH 2: Building HMMs and calculating log-likelihood (contextual anomalies).

Step 1: Train one HMM for each identified time window using a suitable R packages. Provide a rational for choosing your model (parameters).

Step 2: Detect anomalies by comparing the log-likelihood of the training data and the test data (with respect to matching time windows for train and test data).

Step 3: In comparison to a univariate solution, you may also consider a multivariate approach by selecting two or more features for training your HMMs. Recall that selecting feature combinations requires an understanding of their respective correlation coefficients.