CS 780 MMS Class Project

Preliminary Report

Vega Group: Daroc Alden, Samantha Piatt, and Jeremy Walker

April 12, 2018

Motivation

This project is intended to explore opitons for automating a key part of the ongoing Magnetospheric Multiscale Mission - the selection of which detailed datapoints ought to be downloaded from the sattelite. This job is crrently done by a Scientist in the Loop, who must spend time each day evaluating the data observed by the satelite to decide how to spend bandwidth resources.

Automating the selection of interesting data would free up valuable human time on the project. To do this, we have explored several methods for determining what data points might be interested, as detailed below.

Related work

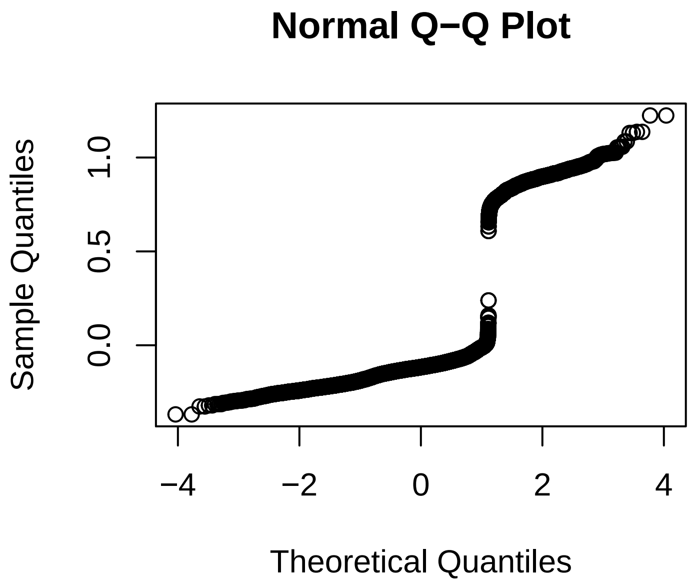
Accuracy of Extrapolation (Time Series) Methods: Results of a Forecasting Competition by S. Makridakit et al. is a definitive comparison of different forecasting methods on different types of and horizons of data. While this is not a forecasting problem per se, the conclusion that they draw is likely appropriate to our problem because of the complex and varied nature of the S.I.T.Ls who decide what data ought to be acquired from the satelite: "[S]tatistically sophisticated methods do not do better than simple methods . . . when there is considerable randomness in the data." (pg. 142 in the Journal of Forecasting, Vol 1, Iss. No. 2)

Rolling Window Selection for Out-of-Sample Forecasting with Time-Varying Parameters by Atsushi Inoue, Lu Jin, and Barbara Rossi reccomends limiting the size of the window of data used for forecasting so that regime changes in the input data don't disrupt the fit of one's model. Unfortunatly, this is not directly applicable to our project either, since we do not have access to all the data up to the present day, but instead only have selected samples.

Time Series Prediction Using Support Vector Machines: A Survey by Jicholas Sapaankevych and Ravi Sankar points out that in the most common application for predicting time series data - the financial sector - a weighted variant of SVMs is often used called C-ascending SVMs, and that these have proven more adaptable, since they emphasize recent or current data over older data.

Evaluation criteria

We found that the data was not identically and independantly distributed, as demonstrated by the exemplary QQ plot below.



In fact the data is a time series. We hypothesise that adding time-delayed factors (i.e. factors which contain values from previous samples) will significantly improve the performance of many machine learning models on the validation set. See below for details.

Methods

We compared our time series models against the basic feature set with no consideration for time, using multiple machine learning methods.

Our first time series model, which we will call time shifting, involves adding n columns to the data that contain each feature shifted n = 5 points in the past. The second time series model, which we willl call time difference, involves adding a column which takes the average differece between the current data point x and the past n = 20 data points.

Random Forests

Using 4 splits (square root of number of features).

	Total.SITL	Found.SITL	Missed.SITL	**Class.Error**	**ERROR**
Test	438	484	46	**0.0195801**	**482.967117**
Train	1897	3363	4	**0.0003215**	**7.590933**
Total	2335	4488	57	**0.0032228**	**107.554431**

Validation

Since, during training, the version of each model using the time difference did significantly better for all models, we will compare those methods' performance on our validation set. Since none of the models perform exceptionally well here, it is likely that they have overfit. We suspect this may be an artifact of the different different SITLs on different weeks, or of our comparativly sparse training data. Regardless, the algorithm's performance on this validation set is a good indicator of its likely performance on the hidden test that you have for us, professor.

Also: Since Logistic Regression did so poorly, we've chosen to omit it here for brevity.

Random Forest

Total.SITL

Found.SITL

Missed.SITL

Class.Error

ERROR

133

134

7

0.0164271

188.922

Recommendations

TODO

Analysis

Since Random Forests did less well on our training data, but better on our validation set, it seems likely that other methods overfit. This can be solved in a few ways, not least of which will be to acquire more data. Overfitting is evidence that our methods are too flexible - so another way to reduce their flexibility would be to use bagging or boosting in combination with another method, or to use cross-validation to choose such parameters as the amount of time-shift to use.