Martingale Based Anomaly Detection – Using Machine Learning to Detect Abnormalities in Time Series Data

Anomaly Detection is the problem of finding patterns in data that do not conform to a model of “normal” behavior. Detecting such deviations from expected behavior in temporal data is important for ensuring the normal operations of systems across multiple domains such as economics, biology, computing, finance, ecology and more. Applications in such domains need the ability to detect abnormal behavior which can be an indication of systems failure or malicious activities, and they need to be able to trigger the appropriate steps towards taking corrective actions. In each case, it is important to characterize what is normal, what is deviant or anomalous and how significant is the anomaly. This characterization is straightforward for systems where the behavior can be specified using simple mathematical models – for example, the output of a Gaussian distribution with known mean and standard deviation. However, most interesting real world systems have complex behavior over time. It is necessary to characterize the normal state of the system by observing data about the system over a period of time when the system is deemed normal by observers and users of that system, and to use this characterization as a baseline to flag anomalous behavior.

In this project, we are trying to develop an Anomaly Detection system that could detect anomalies “On-Line” with adjustable confidence and sensitivity for time-series data. The purpose of this project is not to review the knowledge that you already learn through out the course but to understand how to combine what you learned with something new. The goal of this project is to solve a real world problem and results in something that could be used by users. You will not only do analytic but also get your hands dirty on significant amount of coding.

Milestones:

1. **Project Selection** : Form a team of 3-4 students and select a project from projected list.

2. **Literature Study** : Go through the following recourses for both theoretical background and implementation of the project. Write 1-2 pages of summary for the literature study.
   - Vladimir Vovk, Ilia Nouretdinov, Alex J. Gammerman, “Testing Exchangeability Online”, ICML 2003.

- Shen-Shyang Ho; Wechsler, H., "A Martingale Framework for Detecting Changes in Data Streams by Testing Exchangeability," Pattern Analysis and Machine Intelligence, IEEE Transactions , vol.32, no.12, pp. 2113,2127, Dec. 2010
- Valentina Fedorova, Alex J. Gammerman, Ilia Nouretdinov, Vladimir Vovk, “Plug-in martingales for testing exchangeability on-line”, ICML 2012

3. **Data Exploration and Cleaning**: The primary purpose of this project is to designed a Anomaly Detection module that could be used for generic time series data. Therefore, students could either find or create there own datasets for demo purpose. When finding or creating the datasets, students need to think about :
   - What are different kinds of anomalies that may exist in time series data
   - Are we aiming at detect all of them or a subsets of them.
   - What’re the limitations of the projected algorithm.

   Several sources that may be appropriate from this perspective are :
   - Financial data from yahoo finance: https://finance.yahoo.com
   - Weather data from NOAA: http://www.ncdc.noaa.gov/cdoweb/datasets

4. **Proposal**: Propose the methodology and ideas to be implemented:
   - Decide performance metrics to evaluate the model.
   - Decide what kind of strangeness function supported. Several simple types of strangeness functions are listed below as baseline. Students could also propose their own kind of strangeness function and elaborate their pros/cons:
     - Distance Base
     - Range Percentile
     - OLS positive/negative slope and trend
   - Decide what kind of solution the team would like to present as the result of the project. Students could choose from one or several from the following or propose their own.
     - A python/C/java command line program that take .csv file as input data
     - A program with simple graphic interface that support .csv file
     - A web application hosted somewhere in the cloud that support multiple time series file
   - Propose any possible improvement, either algorithmic or performance wide. These parts are not required by the project but would served further exploration if students are interested.

- What if the input time series data is not uniformly interval? How are we going to interpolate.
- What is the capacity of your implementation? What will happened if the input time series data is too big.
- Are there any way for the system to handle high dimension data instead of just one?