A GUIDE TO OPTIONS IN ADS SETTING FILE

1. Anomaly_level

   This option takes in a number from 0 to 1. It allows the system admin to input the estimated percentage of data that are anomalous, which aids the ADS when making predictions. For example, a value of 0.05 indicates that the system admin estimates that 5% of the input data is anomalous.

2. Maximum stored points

   This option takes in an integer. This is the maximum number of history data that each of the underlying prediction models store in the functions for future prediction references. Prediction accuracy is expected to increase with more data points stored in the models, but at a cost of processing speed and storage space.

3. Excluded Features

   This option takes in a string of features separated by commas. These features will NOT be considered when the ADS is making predictions. For example, the system admin can decide to exclude temporal features such as day/month if the system admin is certain that the data has no significant seasonal trends. This gives the system admin a way to improve the speed of the prediction, but we encourage to leave this option blank if the nature of the data is unknown.

4. Excluded Models

   This option takes in a string of models separated by commas. These models will NOT be considered when the ADS is making predictions. Each underlying model has its own mechanics when it comes to detecting anomalies, some better at detecting certain anomalies than others. If the system admin is certain that a particular model would not be helpful, he or she has the power to exclude it. However, as is with the case of excluded features, we encourage the users to leave this option blank if unsure which models to leave out, because the stacker in the ADS will weigh the result of each model automatically to find the optimal combination for future predictions.

5. Barriers

   This option takes in a string of barrier sets for each dimension of the data. This allows the system admin to specify one or more hard-coded boundaries for each dimension, and the ADS will report 100% of the data within the range of the barriers as anomalies regardless of other settings. This option is particularly useful if the input data has less distinguishable trends and the system admin has a good idea of what is considered "normal". For example, if one of the dimensions is weather temperature, the system admin could set a barrier of [85, inf] to indicate that any temperature reading above 85 is considered anomalous.

6. Prediction delay

    This option takes in an integer. In order for the ADS to accurately deduce anomalies, it requires some initial data. As a result, the ADS not be able to predict for the first few samples. A large prediction delay is likely to give out more accurate predictions from early on, but the tradeoff is the system might have to wait for a period of time before any prediction comes out, especially when the dimension of input data is large.

7. Underlying prediction models

    * REPRESENTATIVE = statistically representative data were chosen to be stored in models
    * RECENT = most recent data were chosen to be stored in models

    BGMM_REPRESENTATIVE and BGMM_RECENT

    Bayesian Gaussian mixture model, or BGMM for short, is a soft clustering model seen unsupervised machine learning for data clustering. Instead of assigning data points to one cluster, BGMM tells the probability that a data point is associated with a specific cluster using prior knowledge about the data set.

    IFOREST_REPRESENTATIVE and IFOREST_RECENT

    Isolation forests is an unsupervised learning algorithm for detecting anomalies in the dataset. It explicitly identifies anomalies instead of profiling normal observations. The general idea is that when building a decision tree from the data set, every split value is chosen by randomly selecting a feature and a value between the minimum and maximum of the selected feature. Since anomalies are rarer than and lie further from the regular data points, they should be separated first and identified closer to the root of the tree.