Luminaire: A hands-off Anomaly Detection Library

version

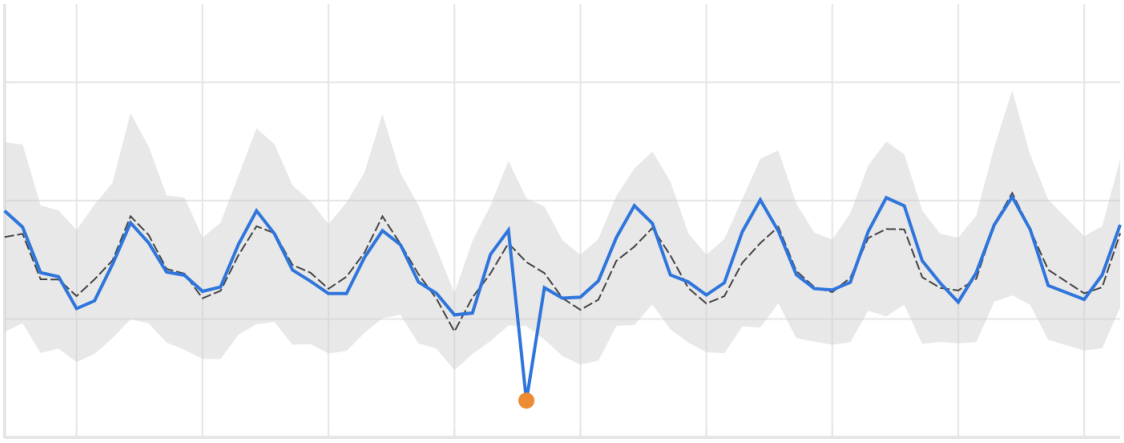
Zillow Group Data Governance team

August 17, 2020

Contents

Luminaire: A hands-off Anomaly Detection Library

Introduction



Luminaire is a python package that provides ML driven solutions for monitoring time series data. Luminaire provides several anomaly detection and forecasting capabilities that incorporate correlational and seasonal patterns in the data over time as well as uncontrollable variations. Specifically, Luminaire is equipped with the following key features:

- **Generic Anomaly Detection:** Luminaire is a generic anomaly detection tool containing several classes of time series models focused toward catching any irregular fluctuations over different kinds of time series data.
- **Fully Automatic:** Luminaire performs optimizations over different sets of hyperparameters and several model classes to pick the optimal model for the time series under consideration. No model configuration is required from the user.
- **Supports Diverse Anomaly Detection Types:** Luminaire supports different detection types:
    - Outlier Detection
    - Data Shift Detection
    - Trend Change Detection
    - Null Data Detection
    - Density comparison for streaming data

Data Exploration and Profiling

Luminaire performs exploratory profiling on the data before progressing to optimization and training. This step provides batch insights about the raw training data on a given time window and also enables automated decisions regarding data pre-processing during the optimization process. These tests and pre-processing steps include:

- Checking for recent data shifts
- Detecting recent trend changes
- Stationarity adjustments
- Imputation of missing data

Outlier Detection

Luminaire generates a model for a given time series based on its recent patterns. Luminaire implements several modeling techniques to learn different variational patterns of the data that ranges from ARIMA, Filtering Models, and Fourier Transform. Luminaire incorporates the global characteristics while learning the local patterns in order to make the learning process robust to any local fluctuations and for faster execution.

Configuration Optimization for Outlier Detection Models

Luminaire combines many techniques under hood to find the optimal model for every time series. Hyperopt is used at its core to optimize over the global hyperparameters for a given time series. In addition, Luminaire identifies whether a time series shows exponential characteristics in terms of its variational patterns, whether holidays have any effects on the time series, and whether the time series shows a long term correlational or Markovian pattern (depending on the last value only).

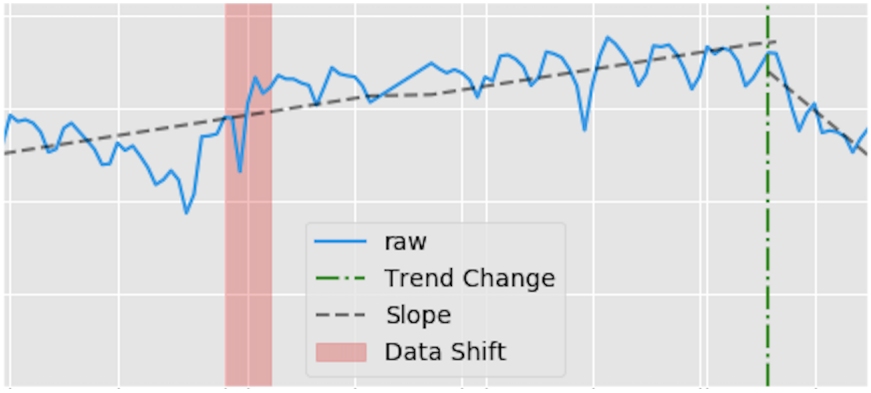
Anomaly Detection for Streaming Data

Luminaire performs anomaly detection over streaming data by comparing the volume density of the incoming data stream with a preset baseline time series window. Luminaire is capable of tracking time series windows over different data frequencies and is autoconfigured to support most typical streaming use cases.

Tutorials

Data Profiling

Luminaire *DataExploration* implements different exploratory data analysis to detect important information from time series data. This method can be used to impute missing data, detect the set of historical trend changes and change points (steady data shifts) which information can later be leveraged downstream in Luminaire outlier detection models.



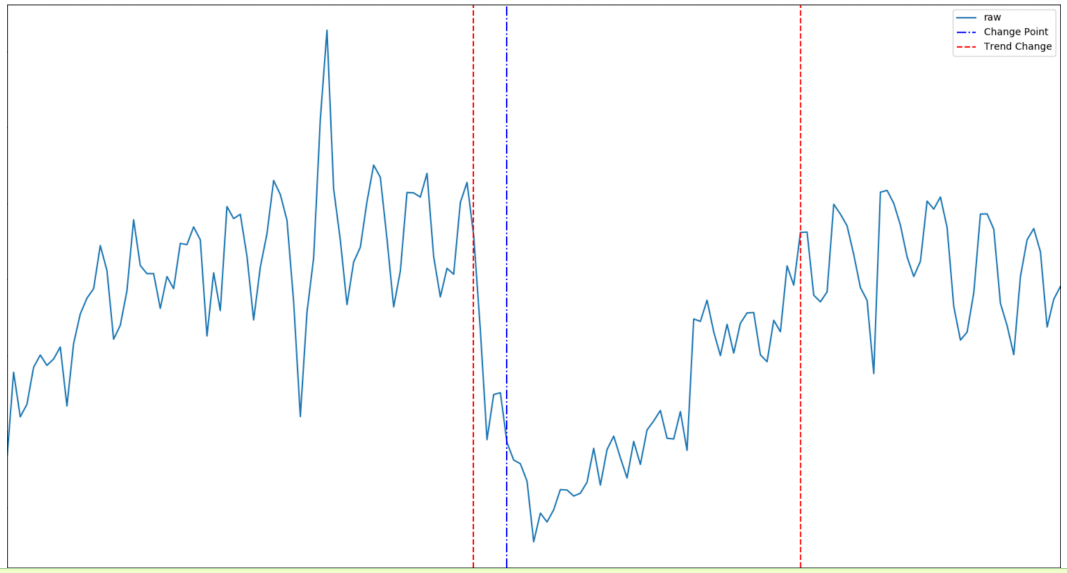
Luminaire data exploration and profiling runs two different workflows. The impute only option in profiling performs imputation for any missing data in the input time series and does not run any profiling to generate insights from the input time series.

```
>>> from luminaire.exploration.data_exploration import DataExploration
>>> data
                raw
index
2020-01-01  1326.0
2020-01-02  1552.0
2020-01-03  1432.0
2020-01-04  1470.0
2020-01-05  1565.0
...            ...
2020-06-03  1934.0
2020-06-04  1873.0
2020-06-05     NaN
2020-06-06  1747.0
2020-06-07  1782.0
>>> de_obj = DataExploration(freq='D')
>>> imputed_data, pre_prc = de_obj.profile(data, impute_only=True)
>>> print(imputed_data)
                   raw
2020-01-01  1326.000000
2020-01-02  1552.000000
```

```
2020-01-03  1432.000000
2020-01-04  1470.000000
2020-01-05  1565.000000
...                 ...
2020-06-03  1934.000000
2020-06-04  1873.000000
2020-06-05  1823.804535
2020-06-06  1747.000000
2020-06-07  1782.000000
>>>print(pre_prc)
None
```

In order to get the data profiling information, the impute only option should be disabled (that is the default option). Disabling the impute only option allows Luminaire to impute missing data along with detecting all the trend changes and the change points in the input time series.

The key utility of Luminaire data profiling is this being a pre-processing step for outlier detection model training. Hence, the user can enable several option to prepare the time series before ingested by the training process. For example, the log transformation option can be enabled for exponential modeling during training. User can also check for the fill rate to constrain the proportion of missing data upto some threshold. Moreover, the pre processed data can also be truncated if there is any change points (data shift) observed.




```
>>> de_obj = DataExploration(freq='D', data_shift_truncate=True, is_log_transformed=True, fi
>>> imputed_data, pre_prc = de_obj.profile(data)
>>> print(pre_prc)
{'success': True, 'trend_change_list': ['2020-04-01 00:00:00'], 'change_point_list': ['2020-
>>> print(imputed_data)
               raw  interpolated
2020-03-16  1371.0      7.224024
2020-03-17  1325.0      7.189922
2020-03-18  1318.0      7.184629
2020-03-19  1270.0      7.147559
2020-03-20  1116.0      7.018401
...            ...           ...
2020-06-03  1934.0      7.567862
2020-06-04  1873.0      7.535830
2020-06-05     NaN      7.610539
2020-06-06  1747.0      7.466227
2020-06-07  1782.0      7.486052
```

Configuration Optimization

Luminaire *HyperparameterOptimization* performs auto-selection of the best data preprocessing configuration and the outlier detection model training configuration with respect to the input time series. This option enables Luminaire to work as a hands-off system where the user only has to provide the input data along with its frequency. This option should be used if the user wants avoid any manual configuration and should be called prior to the data pre-processing and training steps.

```
>>> from luminaire.optimization.hyperparameter_optimization import HyperparameterOptimizatio
>>> print(data)
               raw
index
2020-01-01  1326.0
2020-01-02  1552.0
2020-01-03  1432.0
2020-01-04  1470.0
2020-01-05  1565.0
...            ...
2020-06-03  1934.0
2020-06-04  1873.0
2020-06-05  1674.0
2020-06-06  1747.0
2020-06-07  1782.0
>>> hopt_obj = HyperparameterOptimization(freq='D')
>>> opt_config = hopt_obj.run(data=data)
>>> print(opt_config)
{'LuminaireModel': 'LADStructuralModel', 'data_shift_truncate': 0, 'fill_rate': 0.7423534446
```

Fully Automatic Outlier Detection

Since the optimized configuration contains all the parameters required for data pre-processing and training, this can be used downstream for performing the data pre-processing and training.

```
>>> from luminaire.exploration.data_exploration import DataExploration
>>> de_obj = DataExploration(freq='D', **opt_config)
>>> training_data, pre_prc = de_obj.profile(data)
>>> print(training_data)
               raw  interpolated
2020-01-01  1326.0      7.190676
2020-01-02  1552.0      7.347943
2020-01-03  1432.0      7.267525
2020-01-04  1470.0      7.293697
2020-01-05  1565.0      7.356279
...            ...           ...
2020-06-03  1934.0      7.567862
2020-06-04  1873.0      7.535830
2020-06-05  1674.0      7.423568
2020-06-06  1747.0      7.466227
2020-06-07  1782.0      7.486052
```

The above piece of code makes the data ready to be ingested for training. The only step left before training is to extract the luminaire outlier detection model object for the optimized configuration.

```
>>> model_class_name = opt_config['LuminaireModel']
>>> module = __import__('luminaire.model', fromlist=[''])
>>> model_class = getattr(module, model_class_name)
>>> print(model_class)
<class 'luminaire_models.model.lad_structural.LADStructuralModel'>
```

Since, we have to optimal model class along with other optimal configurations, we can run training as follows:

```
>>> model_object = model_class(hyper_params=opt_config, freq='D')
>>> success, model_date, trained_model = model_object.train(data=training_data, **pre_prc)
```

```
>>> print(success, model_date, trained_model)
(True, '2020-06-07 00:00:00', <luminaire_models.model.lad_structural.LADStructuralModel obje
```

This trained model is now ready to be used for scoring future data points.

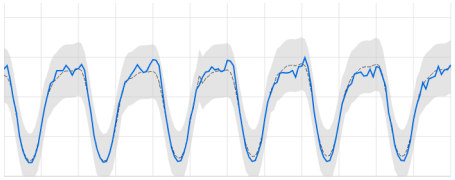
```
>>> trained_model.score(2000, '2020-06-08')
{'Success': True, 'IsLogTransformed': 1, 'LogTransformedAdjustedActual': 7.601402334583733,
```

Outlier Detection

Luminaire can detect outliers in time series data by modeling the predictive and the variational patterns of a time series trajectory. Luminaire is capable of tracking outliers for any time series data by applying two specific modeling capabilities:

- **Structural Model**: This technique is suitable for time series datasets that show periodic patterns and contains good predictive signals through temporal correlations.
- **Filtering Model**: This technique is suitable for noisy time series datasets that contains almost no predictive signals from the periodic or temporal correlation signals.

Anomaly Detection using Structural Model



Luminaire provides the full capability to have user-specified configuration for structural modeling. Under the hood, Luminaire implements a linear or an exponential model allowing multiple user specified auto regressive and moving average components to track any temporal correlational patterns. Fourier transformation can also be applied under the hood if the data shows strong seasonality or periodic patterns. As external structural information, Luminaire allows holidays to be added as external exogenous features (currently supported for daily data only) inside the structural model.

```
>>> from luminaire.model.lad_structural import LADStructuralModel
>>> hyper_params = {"include_holidays_exog": True, "is_log_transformed": False, "max_ft_freq
>>> lad_struct_obj = LADStructuralModel(hyper_params=hyper, freq='D')
>>> print(lad_struct_obj)
<luminaire_models.model.lad_structural.LADStructuralModel object at 0x7fc91882bb38>
```

Luminaire allows some data-specific information to be added during the training process of the structural model through *preprocessing_parameters*. The *preprocessing_parameters* can either be specified by the user if the data-specific information is available through external sources OR can be obtained using *Luminaire DataExploration*. The data-specific information includes a list of trend changes, change points (data shifts), and start and end of the input time series.

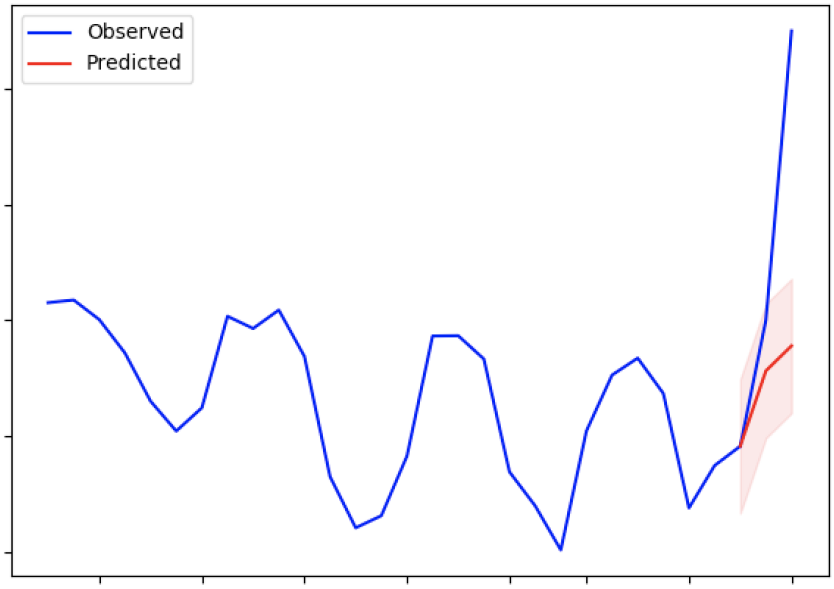
```
>>> from luminaire.exploration.data_exploration import DataExploration
>>> de_obj = DataExploration(freq='D', data_shift_truncate=False, is_log_transformed=True, f
>>> data, pre_prc = de_obj.profile(data)
>>> print(pre_prc)
{'success': True, 'trend_change_list': ['2020-04-01 00:00:00'], 'change_point_list': ['2020-
'is_log_transformed': 1, 'min_ts_mean': None, 'ts_start': '2020-01-01 00:00:00', 'ts_end': '
```

These *preprocessing_parameters* are used for training the structural model.

```
>>> success, model_date, model = lad_struct_obj.train(data=data, **pre_prc)
>>> print(success, model_date, model)
(True, '2020-06-07 00:00:00', <luminaire_models.model.lad_structural.LADStructuralModel obje
```

The trained model works as a data-driven source of truth to evaluate any future time series values to be monitored. The *score* method is used to check whether new data points are anomalous.

```
>>> model.score(2000, '2020-06-08')
{'Success': True, 'IsLogTransformed': 1, 'LogTransformedAdjustedActual': 7.601402334583733,
>>> model.score(2500, '2020-06-09')
{'Success': True, 'IsLogTransformed': 1, 'LogTransformedAdjustedActual': 7.824445930877619,
```



The scoring function outputs several fields. The key to identifying whether a data point has been detected as an anomaly is the *AnomalyProbability* field (for anomalous fluctuations in either direction) and *DownAnomalyProbability*, *UpAnomalyProbability* for one-sided fluctuations that are lower or higher than expected, respectively. The user can set any anomaly threshold to identify whether a pont is an anomaly or not. From the above example, by setting the anomaly threshold at *0.99* for both sided fluctuations, we can see the the the value corresponding to *2020-06-08* is non anomalous whereas the value for *2020-06-09* is anomalous. Luminaire also has its own pre-specified thresholds at *0.9* and at *0.999* for identifying mild an extreme anomalies (see the keys *IsAnomaly* and *IsAnomalyExtreme*).

Luminaire generates a *ModelFreshness* score to identify how fresh the model is (i.e. what is the difference between the scoring data date and the model date). This freshness scores varies between 0 to 1 and the model object expires whenever the freshness score exceeds the value 1.

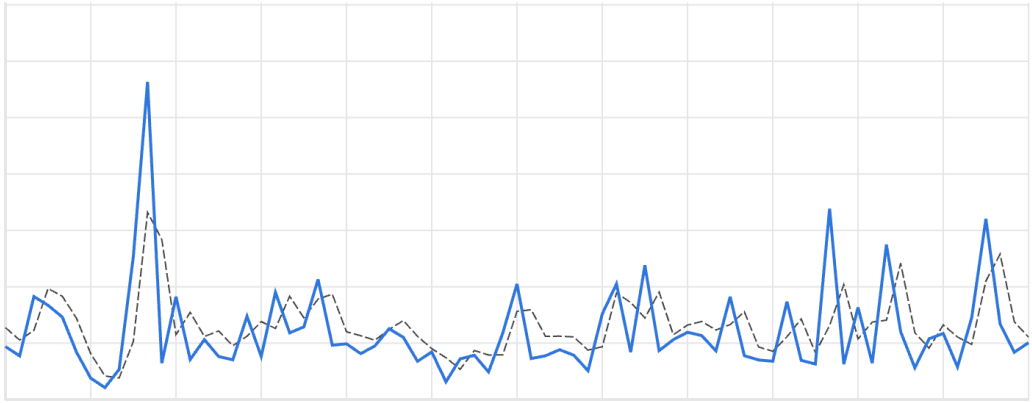
Forecasting

Since the anomaly detection process through structural modeling depends on quantifying the predictive and the variational patterns of the underlying data, Luminaire also outputs several forecasting metrics such as *Prediction*, *StdErr*, *CILower* and *CIUpper* that can be used for time series forecasting use cases.

Note

The *ConfLevel* in the scoring output corresponds to the generated confidence intervals and to the *IsAnomaly* flag

Anomaly Detection using Filtering Model



Luminaire allows monitoring noisy and not too well behaved time series data by tracking the residual process from a filtering model. This model should not be used for predictive purposes but can be used to measure variational patterns and irregular fluctuations.

Filtering requires very minimal specification in terms of configurations. The user needs to only configure whether to implement a linear or exponential model.

```
>>> from luminaire.model.lad_filtering import LADFilteringModel
>>> hyper = {"is_log_transformed": 1}
>>> lad_filter_obj = LADFilteringModel(hyper_params=hyper, freq='D')
>>> print(lad_filter_obj)
<luminaire_models.model.lad_filtering.LADFilteringModel object at 0x7fd2b1832dd8>
```

Similar to the structural model, the user can specify the *preprocessing_parameters* (see lad structural modeling tutorial for further information). These *preprocessing_parameters* are required to train the Luminaire filtering model.

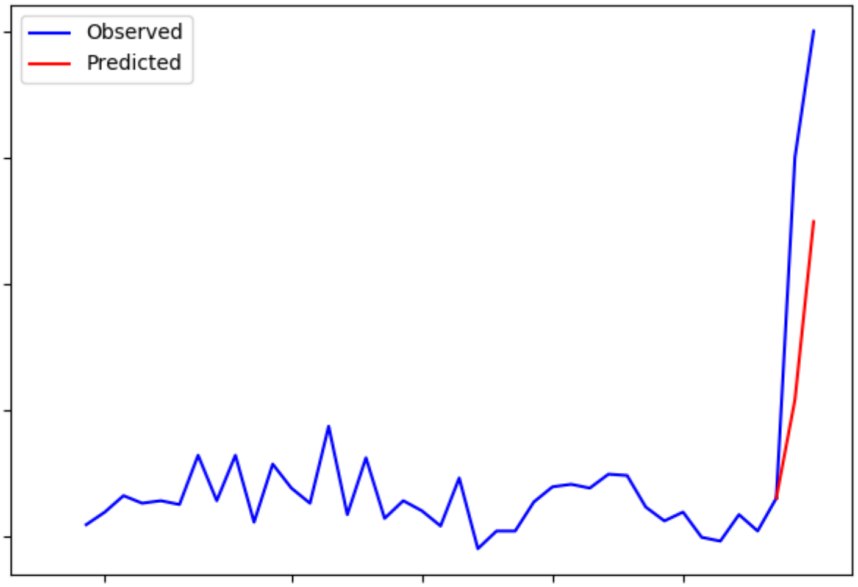
```
>>> success, model_date, model = lad_filter_obj.train(data=data, **pre_prc)
>>> print(success, model_date, model)
(True, '2019-08-27 00:00:00', <luminaire_models.model.lad_filtering.LADFilteringModel object
```

Similar to the structural model, this trained filtering model can be used to score any future time series values. Moreover, the filtering model updates some components of the model object every time it scores to keep the variational information updated.

```
>>> scores, model_update = model.score(400, '2019-08-28')
>>> print(scores, model_update)
({'Success': True, 'AdjustedActual': 1.4535283491638031, 'ConfLevel': 90.0, 'Prediction': 20
```

The trained *model* can only be used to score the next innovation after the training. To score any futher points in the future, the iterative *model_update* needs to be used.

```
>>> scores_2, model_update_2 = model_update.score(500, '2019-08-29')
>>> print(scores_2, model_update_2)
({'Success': True, 'AdjustedActual': -0.591849553174421, 'ConfLevel': 90.0, 'Prediction': 34
```



Note

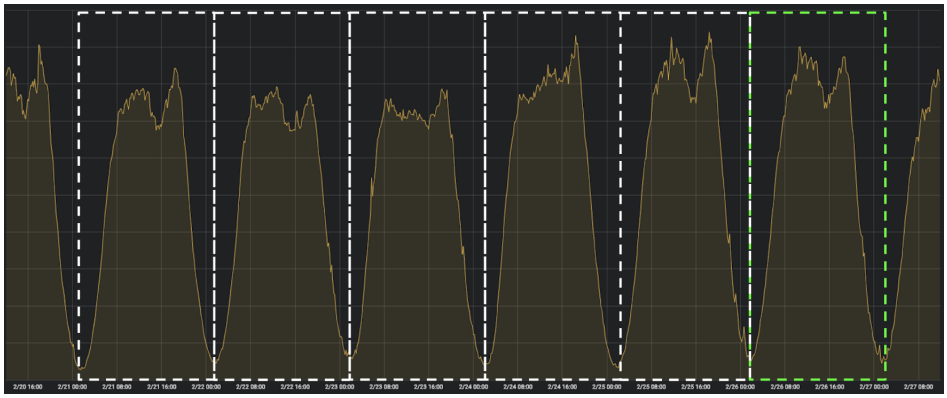
Prediction for the filtering model is a posterior prediction, which means the prediction is made after observing the data to score. See kalman_filter for more information.

Note

It is important to note that the model update process during scoring only updates a small portion of the model component. It is a good practice to train the model over some schedule to acheive the best performance.

Anomaly Detection for Streaming data

Luminaire *WindowDensityModel* implements the idea of monitoring data over comparable windows instead of tracking individual data points as outliers. This is a useful approach for tracking anomalies over high frequency data, which tends to show a higher level of noise. Hence, tracking anomalies over streaming data essesntially means tracking sustained fluctuations.



Although *WindowDensityModel* is designed to track anomalies over streaming data, it can be used to track any sustained fluctuations over a window for any frequency. This detection type is suggested for up to hourly data frequency.

Anomaly Detection: Pre-Configured Settings

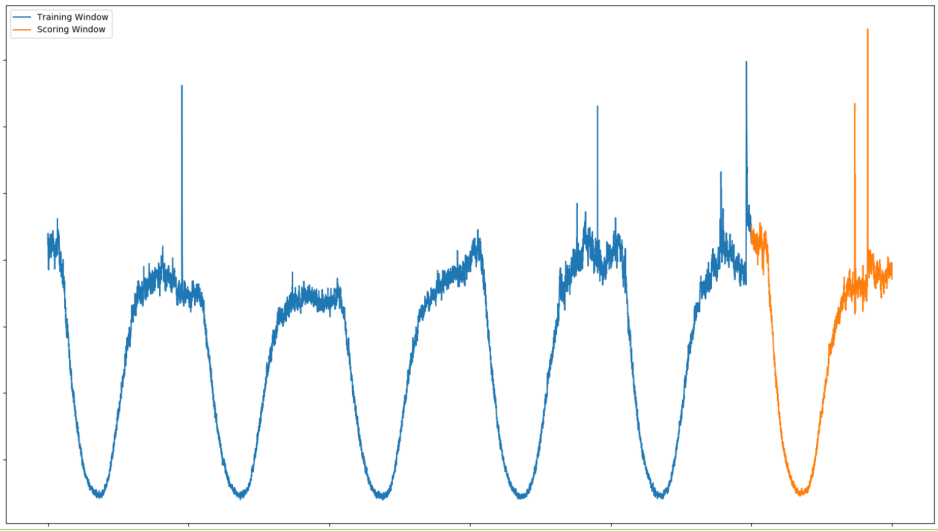
Luminaire provides the capability to configure model parameters based on the frequency that the data has been observed and the methods that can be applied (please refer to the Window density Model user guide for detailed configuration options). Luminaire settings for the window density model are already pre-configured for some typical pandas frequency types and settings for any other frequency types should be configured manually (see the user guide for more information).

```
>>> from luminaire.model.window_density import WindowDensityHyperParams, WindowDensityModel
>>> print(data)
                        raw  interpolated
index
2020-05-25 00:00:00  10585.0       10585.0
2020-05-25 00:01:00  10996.0       10996.0
2020-05-25 00:02:00  10466.0       10466.0
2020-05-25 00:03:00  10064.0       10064.0
2020-05-25 00:04:00  10221.0       10221.0
...                      ...           ...
2020-06-16 23:55:00  11356.0       11356.0
2020-06-16 23:56:00  10852.0       10852.0
2020-06-16 23:57:00  11114.0       11114.0
2020-06-16 23:58:00  10663.0       10663.0
2020-06-16 23:59:00  11034.0       11034.0
>>> hyper_params = WindowDensityHyperParams(freq='M').params
>>> wdm_obj = WindowDensityModel(hyper_params=hyper_params)
>>> success, model = wdm_obj.train(data=data)
>>> print(success, model)
(True, <luminaire_models.model.window_density.WindowDensityModel object at 0x7f8cda42dcc0>)
```

The model object contains the data density structure over a pre-specified window, given the frequency. Lumianire sets the following defaults for some typical pandas frequencies (any custom requirements can be updated in the hyperparameter object instance):

- ‘S’: Hourly windows
- ‘M’: Daily windows
- ‘QM’: Weekly windows
- ‘H’: 12 hours windows
- ‘D’: 10 days windows
- ‘custom’: User specified windows

In order to score a new window innovation given the trained model object, we have to provide a equal sized window that represents a similar time interval. For example, if each of the windows in the training data represents a 24 hour window between 9 AM to 8:59:59 AM for last few days, the scoring data should represent the same interval of a different day and should have the same window size.




```
>>> scoring_data
                        raw interpolated
index
2020-06-17 00:00:00  1121.0       1121.0
2020-06-17 00:01:00  1091.0       1091.0
2020-06-17 00:02:00  1063.0       1063.0
2020-06-17 00:03:00  1085.0       1085.0
2020-06-17 00:04:00  1063.0       1063.0
...                     ...          ...
2020-06-17 23:55:00   968.0        968.0
2020-06-17 23:56:00   995.0        995.0
2020-06-17 23:57:00   963.0        963.0
2020-06-17 23:58:00   968.0        968.0
2020-06-17 23:59:00   920.0        920.0
>>> scores = model.score(scoring_data)
>>> print(scores)
{'Success': True, 'ConfLevel': 99.9, 'IsAnomaly': False, 'AnomalyProbability': 0.69567457348
```

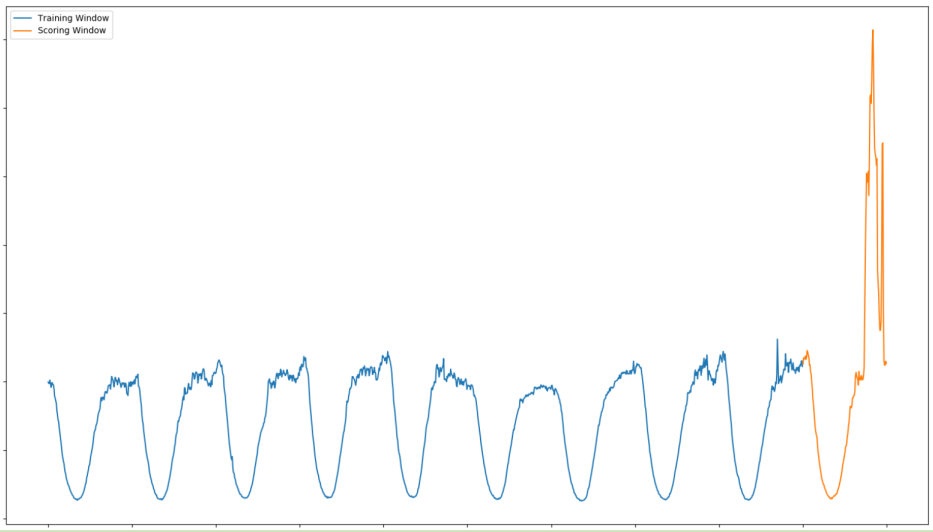
Anomaly Detection: Manual Configuration

There are several options in the *WindowDensityHyperParams* class that can be manually configured. The configuration should be selected mostly based on the frequency that the data has been observed.

```
>>> from luminaire.model.window_density import WindowDensityHyperParams, WindowDensityModel
>>> print(data)
                            raw interpolated
index
2020-05-20 00:03:00  6393.451190  6393.451190
2020-05-20 00:13:00  6491.426190  6491.426190
2020-05-20 00:23:00  6770.469444  6770.469444
2020-05-20 00:33:00  6490.798810  6490.798810
2020-05-20 00:43:00  6273.786508  6273.786508
...                          ...          ...
2020-06-09 23:13:00  5619.341270  5619.341270
2020-06-09 23:23:00  5573.001190  5573.001190
2020-06-09 23:33:00  5745.400000  5745.400000
2020-06-09 23:43:00  5761.355556  5761.355556
2020-06-09 23:53:00  5558.577778  5558.577778
>>>hyper_params = WindowDensityHyperParams(freq='custom',
                                           detection_method='kldiv',
                                           baseline_type="last_window",
                                           min_window_length=6*12,
                                           max_window_length=6*24*84,
                                           window_length=6*24,
```

```
                                            ma_window_length=24,
                                            ).params
>>> wdm_obj = WindowDensityModel(hyper_params=hyper_params)
>>> success, model = wdm_obj.train(data=data)
>>> print(success, model)
(True, <luminaire_models.model.window_density.WindowDensityModel object at 0x7f8d5f1a6940>)
```

The trained model object can be used to score data representing the same interval from a different day and having the same window size.




```
>>> scoring_data
                            raw interpolated
index
2020-06-10 00:00:00  5532.556746  5532.556746
2020-06-10 00:10:00  5640.711905  5640.711905
2020-06-10 00:20:00  5880.368254  5880.368254
2020-06-10 00:30:00  5842.397222  5842.397222
2020-06-10 00:40:00  5827.231746  5827.231746
...                          ...          ...
2020-06-10 23:10:00  7210.905952  7210.905952
2020-06-10 23:20:00  5739.459524  5739.459524
2020-06-10 23:30:00  5590.413889  5590.413889
2020-06-10 23:40:00  5608.291270  5608.291270
2020-06-10 23:50:00  5753.794444  5753.794444
>>> scores = model.score(scoring_data)
>>> print(scores)
{'Success': True, 'ConfLevel': 99.9, 'IsAnomaly': True, 'AnomalyProbability': 0.999999985183
```

User Guide

Luminaire Outlier Detection Models: Structural Modeling

Luminaire Outlier Detection Models: Factoring holidays as exogenous

Luminaire Outlier Detection Models: Kalman Filter

Data Exploration and Profiling

Luminaire Configuration Optimization

Luminaire Streaming Anomaly Detection Models: Window Density Model

Indices and tables

- `genindex`
- `modindex`
- `search`