Notes on Anomaly Detection

Sameer Kesava PhD

Contents

Unsupervised Learning Algorithms

For more information, please see **Deep Learning for Anomaly Detection: A Survey**; http://arxiv.org/abs/1901.03407.

1.1 MeanShift Clustering

- Distance-based clustering
- Has the ability to detect outliers.

1.2 DBSCAN

Density-based spatial clustering of applications with Noise

- Hyperparameters:
  1. min_samples: controls how tolerant the algorithm is towards noise. Starting value: 2*dimension.
  2. ϵ: crucial parameter; controls the local neighborhood of points. A starting value can be chosen using elbow/knee point in nearest neighbor plot.
  3. ϵ-metric: euclidean, minkowski, etc.
- Has the ability to detect outliers: A sample that is not a core sample and is at least ϵ in distance from any sample is considered an outlier.
- Works well for non-linear data.
- Affected by the curse of dimensionality.
- OPTICS is a variant of DBSCAN, does not require ϵ to be set.

1.3 Local Outlier Factor

See sklearn

Supervised Learning Algorithms

2.1 Univariate Data

- Boxplot
- Grubbs test
- RANSAC algorithm for linear regression
- Studentized residuals and leverage points. Easy to plot for univariate data.

2.2 Multivariate Data

2.3 Random Cut Forest

From Amazon SageMaker

2.4 XGBoost

Highly popular classifier and regressor.

- Gradient boosting method
- Absolute loss and Huber loss more robust to outliers.
- Hyperparameters
  1. Max_depth
  2. Colsample_bytree
  3. Eta
  4. train-test split: 60-40/70-30/80-20.

2.5 Isolation Forest

Improving the Accuracy

3.1 Hyperparameter Tuning

- Hyperparameter optimization based on Gaussian Process Regression and Bayesian Optimization
- keras tuner in keras
- GridSearchCV or RandomSearchCV in scikit-learn
- RandomSearch can be used as the baseline against which optimization algorithms can be evaluated.

Bibliography

[1] Pankaj Malhotra et al., Long Short Term Memory Networks for Anomaly Detection in Time Series, 2015.