Notes on Anomaly Detection

Sameer Kesava PhD

Contents

1 Supervised Learning Algorithms **1**
1.1 Univariate Data 1
1.2 Multivariate Data 1
1.3 Random Cut Forest 1
1.4 XGBoost 1

2 Unsupervised Learning Algorithms **3**
2.1 MeanShift Clustering 3
2.2 DBSCAN 3

3 Improving the Accuracy **4**
3.1 Hyperparameter Tuning 4

Supervised Learning Algorithms

1.1 Univariate Data

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

1.2 Multivariate Data

1.3 Random Cut Forest

From Amazon SageMaker

1.4 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.

Unsupervised Learning Algorithms

2.1 MeanShift Clustering

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

2.2 DBSCAN

Density-based spatial clustering of applications with Noise

- Hyperparameters: min_samples, ϵ and ϵ-metric.
- Has the ability to detect outliers.
- Works well for non-linear data.
- Affected by the curse of dimensionality.

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.