Anomaly Detection

Jean Paul ISHIMWE

African Institute for Mathematical Sciences-Ghana
AMMI program

29/11/2019

Outlines

AIMS African Institute for Mathematical Sciences GHANA

Introduction

Often times, we normally need to find abnormal and unusual values in our data set. Having outlier can really throw a wrench in training a machine learning model. Thus, this process of finding out abnormality and strange observation from the data set is referred to as **anomaly detection**. What purpose does this really serve?

Cont'Intro

Identifying these data points can serve multiple purposes, such as

1. Removing outliers in a training set before fitting a machine learning model.
2. Analyzing a set of observations to identify whether something is wrong with the process. For example, are there any anomalous server logs that *may* indicate a security breach.

Three settings

1. **Supervised**
   Training data labeled with *normal* or *anomaly*
2. **Clean**
   Training data are all *normal*, test data contaminated with *anomaly* points
3. **Unsupervised**
   Training data consist of mixture of *normal* and *anomaly* points

Methods for anomaly detection

Distinguishing whether something is abnormal or not, falls into two divisions, **outlier** or **novelty detection**.

- **Outlier Detection:** Is the process of identifying observations that deviate substantially a lot from the rest of data points.

  Outlier detection models are trained with *unclean* datasets. And the model learns how much a point can deviate to be classified as an outlier.

Cont'MAD

- **Novelty Detection:** Meant the process of identifying novel points when a model is trained on an *unpolluted* data set with outliers.

  i) The model learns a boundary, or boundaries, that encompasses all normal points.
  ii) Any point that reside outside of these boundaries is classified as novel

Note: This classification does not mean is a supervised learning algorithm but we want to flag the points as in liers or outliers.

Algorithms for Anomaly detection

1. Density-Based Approaches
   - Robust kernel Density Estimation
   - Ensemble Gaussian Mixture Model
2. Quantile-Based Methods
   - One class SVM
   - Support Vector Data Description
3. Neighbor-Based Methods
   - Local Outlier Factor
   - kNN Angle-Based Outlier Detector
4. Projection-Based Methods
   - Isolation Forest (iForest)
   - Lightweight Online Detector of Anomaly

AIMS African Institute for Mathematical Sciences GHANA

Con't Algorithms

Anomaly detection algorithms are unsupervised learners. From `scikit-learn`, we consider **one-class SVM** and **Isolation forest**.

1) **Isolation forest(ISOF)**: Is the outlier detection algorithm that is based on growing a tree[3].

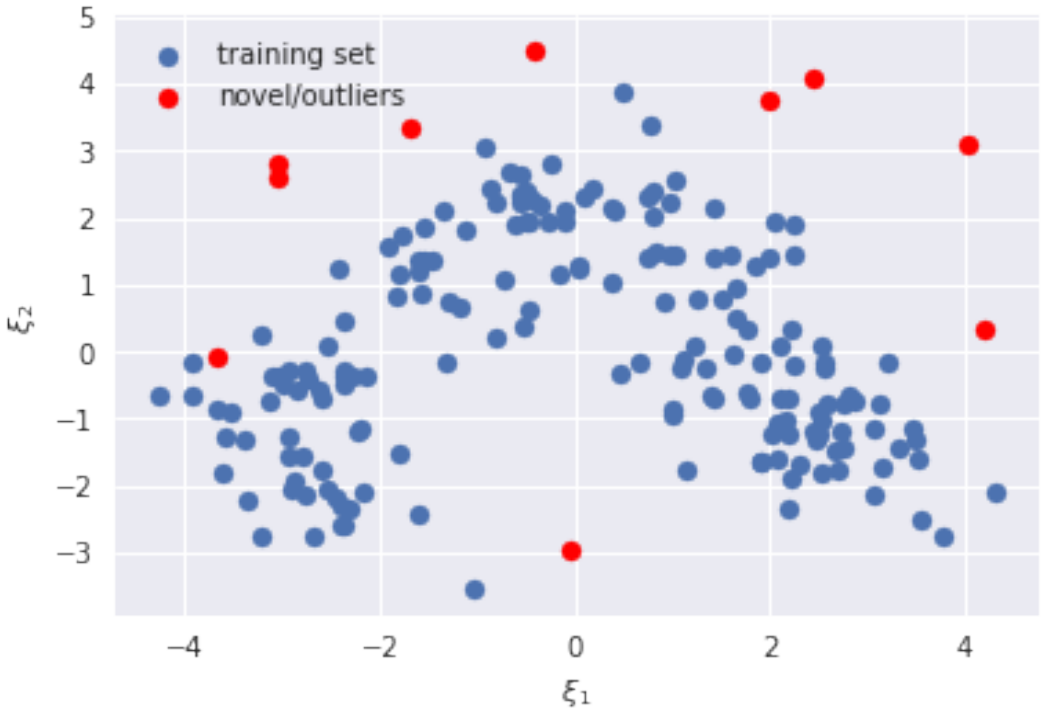


Figure: Training set and novel/outliers

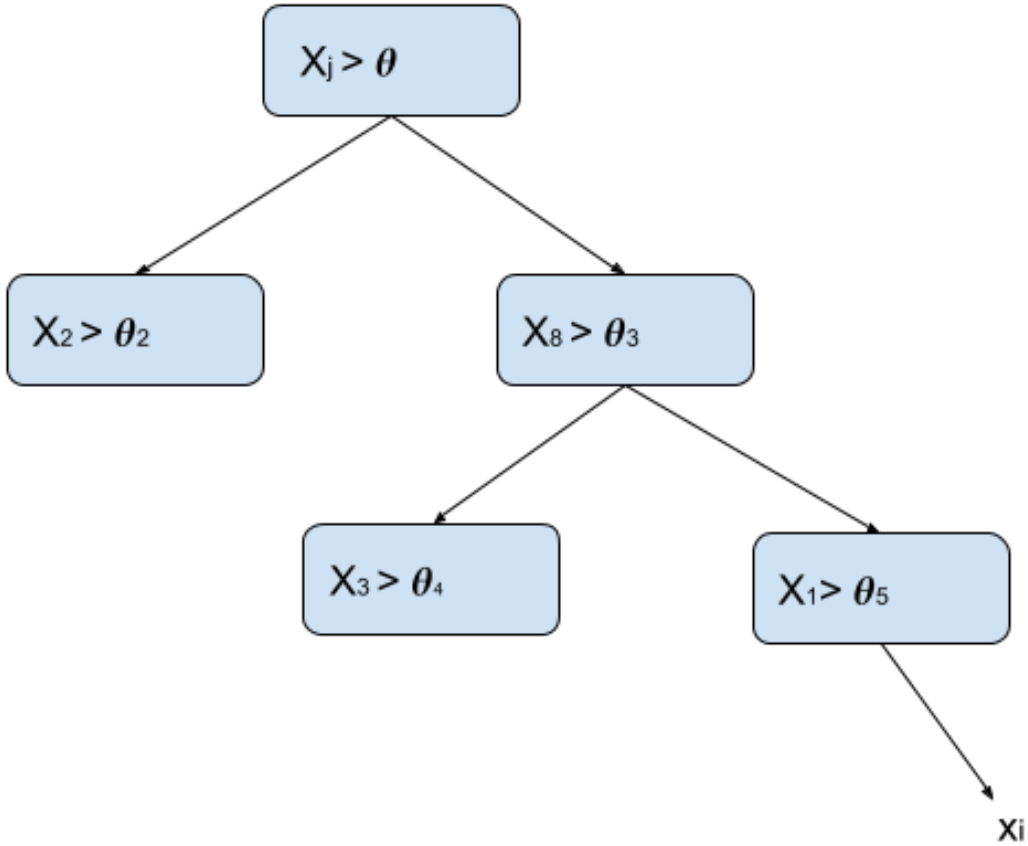
Con't Algorithms

Given a dataset $D = \{x_1, x_2 \cdots x_n\}$ of n instances,

1. Construct a fully random binary tree
   - choose attribute j at random
   - choose spiting threshold θ uniform from $[min(x_j), max(x_j)]$
   - until every data from root node is in its own leaf(1 observation)
   - Let $d(x_i)$ be depth of point x_i
2. repeat 100 times
   - Let $\overline{d}(x_i)$ be average depth of point x_i
   - score(xi) $= 2^{-\frac{\overline{d}(x_i)}{r(xi)}}$ $=$ mean path length $-$ offset, where $r(xi)$ is the expected depth

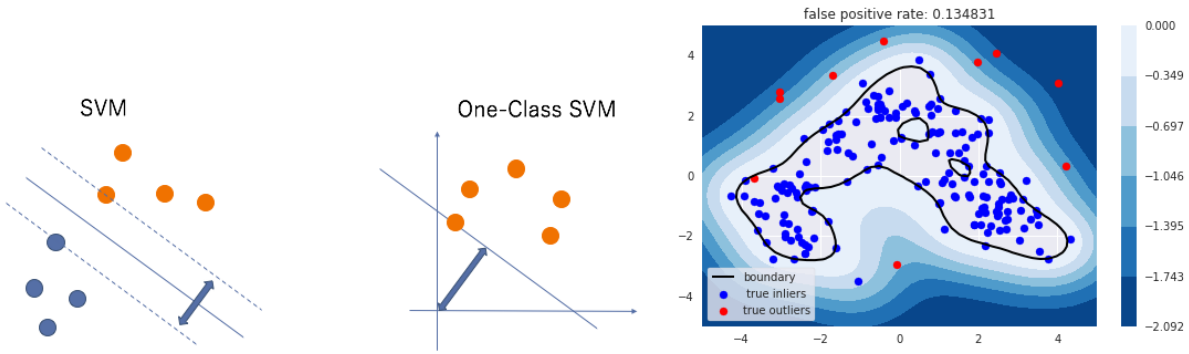
The smaller the average of path length, the more likely the point x is an outlier.

This score is computed by `decision_function` of ISOF to assign observation as an outlier or normal.



Cont'Algorithms for AD

2) **One-class SVM**: The SVM classifier can be tweaked to serve novelty detection applications, referred to as one-class support vector machines. Where all points in the training set belong to the same class, hence the name of the algorithm. Because all training points are in the same class, it is assumed that there are no outliers.

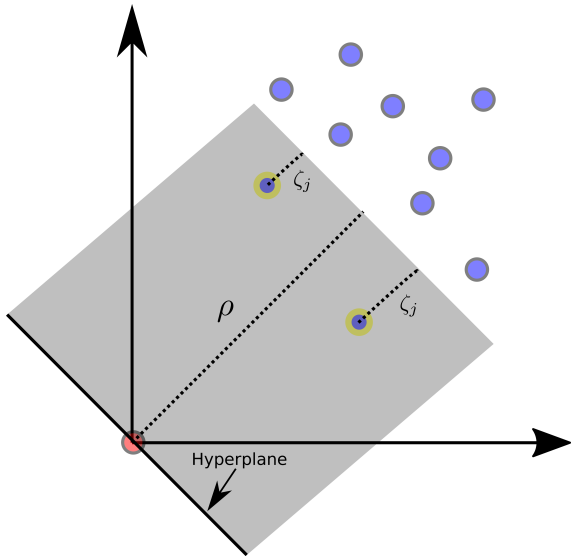


(a) (b)

Figure: One SVM Classifier: (a) Boundary (b) Contours boundaries

Cont'Algorithms for AD

The governing equation and constraints for 1-svm from fig.2b are [1]:



$$\min_{\beta,\zeta,\rho} \frac{1}{2}\|\beta\|^2 + \frac{1}{\nu n}\sum_{j=1}^{n}\zeta_j - \rho$$

$$\text{subject to } \begin{cases} h(x_j.)\cdot\beta \geq \rho - \zeta_j \\ \zeta_j \geq 0, \end{cases}$$

AIMS African Institute for Mathematical Sciences GHANA

where $h(x_j)$ is a kernel function,

ρ is distance from the origin to the hyperplane, and

ν is the probability of finding a new, but regular, observation outside the frontier.

Notice how the constraint is forcing points to be at least ρ away from the margin, lest it incurs a margin violation as ζ must be set large enough to satisfy the inequality.

z-score

The statistical method, **z-score** is a relative measure of how far away a value is from the mean, normalized by the standard deviation.

$$z = \frac{x - \mu}{\sigma},$$

where μ and σ are the mean and standard deviation of distribution. The larger the magnitude of the z-score, the lower the probability of observing the value. Exact percentages can only be known if we know the distribution. When the distribution is a normal or Gaussian distribution given by

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1}$$

68% of the values will reside within $z = \pm 1$. 95% and 99.7% of the values will reside in an interval of $z = \pm 2$ and $z = \pm 3$, respectively.
The general idea of the z-score is the greater is its magnitude the more anomalous the observation is valid. Any observation above the cutoff is flagged as an outlier otherwise is an inlier [2].

Density estimation

Given an unlabeled training data set $M = \{x^1, \cdots, x^m\}$ and each example $x \in \mathbb{R}^n$, where n is the number of features. If assumed that the features follow the normal distribution,

$$\mathcal{X}_i \sim \left(\mu_i, \sigma_i^2\right)$$

We can now model the anomaly detection model $P(X)$ from this dataset and evaluate the highest and low probabilities.

$$\mathcal{P}(X) = P(x_1; \mu_1, \sigma_1^2) \cdot P(x_2; \mu_2, \sigma_1^2) \cdots P(x_n; \mu_n, \sigma_1^n) \tag{2}$$

AIMS African Institute for Mathematical Sciences GHANA

Detection Algorithm

1. Choose features x_i that might be an indicative of anomalous example
2. Fit parameters $\mu_1, \mu_2 \cdots \mu_n$ and $\sigma^1, \sigma^2, \cdots \sigma^n$

$$\begin{cases} \mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \\ \sigma_j^2 = \frac{1}{m} \sum_{i=1}^{m} \left(x_j^{(i)} - \mu_j \right)^2 \end{cases}$$

3. Given a new example x, compute $\mathcal{P}(X)$, by fitting the data and the estimates from step 2,

$$\mathcal{P}(x) = \prod_{i=1}^{n} P(x_j; \mu_j, \sigma_j^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left(-\frac{(x_j - \mu_j)^2}{2\sigma_j^2} \right) \quad (3)$$

Algorithm Evaluation

If the probability from equation (3) is calculated, The new observation is anomalous:

$$\begin{cases} if \quad \mathcal{P}(x) < \epsilon, & Anomalous \\ \quad\quad \mathcal{P}(x) > \epsilon, & Normal \end{cases}$$

The ϵ is the desired threshold to flag an unusual observation. The Anomaly detection algorithms may sometimes be used for datasets with rarely positive label. If this is the case, cross-validation is performed on few observation samples of anomalous label and evaluate the model (3) and choosing maximized **F-1 score** from tuning ϵ.

AIMS African Institute for Mathematical Sciences GHANA

Conclusion

- The anomaly detectors are robust for isolating outliers
- They explicitly isolates anomalies instead of profiling normal instances

References

Chih-Chung Chang and Chih-Jen Lin
LIBSVM: A Library for Support Vector Machines
March 4, 2013.

WorldQuant University
Anomaly Detection
2019.

Sklearn documentation
Novelty and Outlier Detection
https://scikit-learn.org/stable/modules/outlier_detection.html.

Me-Daa-sse:

Thank you!.