Business Analytics II

Lecture 3

Anomaly Detection with Python

Pedro Hernandez Serrano
p.hernandezserrano@maastrichtuniversity.nl

April 10, 2018

BISS Institute

Today's session

- Anomaly Detection in Python.
  - What are Anomalies??
  - Use cases.
  - Point anomalies.
  - Collective anomalies
  - K-means algorithmn
- Hands-on during the session on Jupyter Notebooks
  - Construct Turicreate Methods.
  - Treating on different datasets.

Anomaly Detection with Python

Anomaly Detection

What is Anomaly Detection?

Refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains.

- Point Anomalies
- Collective Anomalies
- Contextual Anomalies

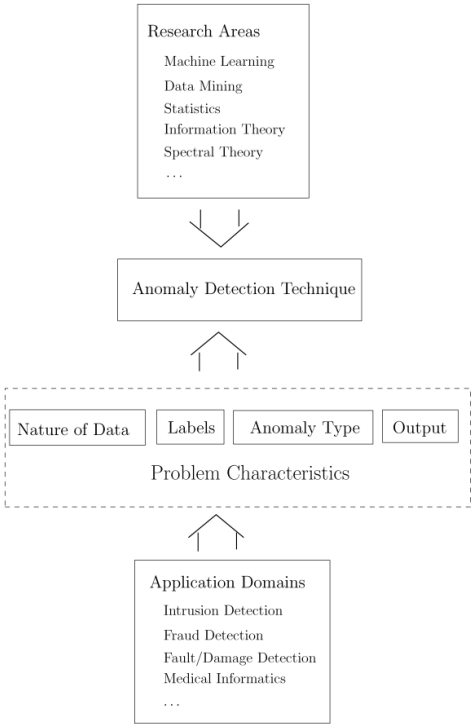
Challenges

A straightforward anomaly detection approach, therefore, is to define a region representing normal behavior and declare any observation in the data that does not belong to this normal region as an anomaly. But several factors make this approach very challenging:

- Defining a normal region that encompasses every behavior is difficult.
- The boundary between normal and anomalous behavior is not precise.
- When anomalies are the result of malicious actions, the malicious adversaries often adapt themselves to make the anomalous observations appear normal, thereby making the task of defining normal behavior more difficult.
- In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.

Challenges

- The exact notion of an anomaly is different for different application domains (e.g. medical biometrics vs stock market).
- Availability of labeled data for training/validation of models used by anomaly detection techniques is usually a major issue.
- Often the data contains noise that tends to be similar to the actual anomalies and hence is difficult to distinguish and remove.
- In many domains normal behavior keeps evolving and a current notion of normal behavior might not be sufficiently representative in the future.



Nature of Input Data

- The nature of attributes determines the applicability of anomaly detection techniques.
- Identify the minimum aggregate level of the anomaly class (transaction, record, measure, etc.)
- Trying pairwise distance between features might be provided in the form of a distance or similarity matrix.
- Always take in consideration the scale and data type on every feature.

Point Anomalies

Point Anomaly (Outlier)

Also known as Outlier, occurs when an individual data instance can be considered as anomalous with respect to the rest of data, then the instance is termed a point anomaly.

An outlying observation is one that appears to **deviate** markedly from other members of the sample in which it occurs.

Note: An outlying observation may be merely an extreme manifestation of the random variability inherent in the data.

Point Anomalies

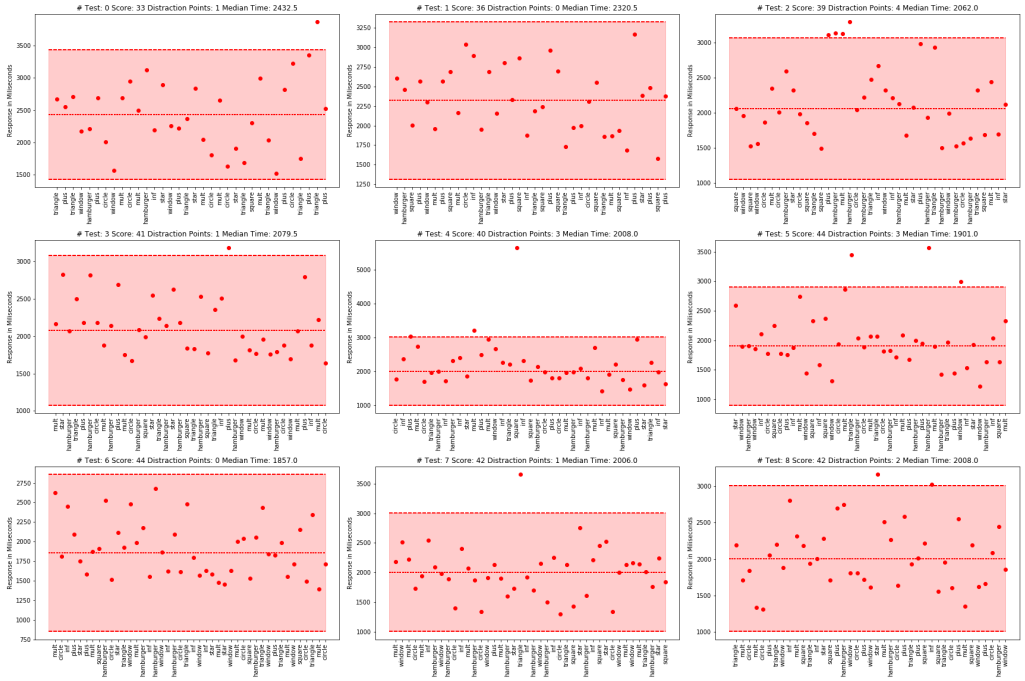
Point Anomaly (Outlier)

In any other case:

- Count the frequency.
- Label observation(s).
- Reject observation(s).
- Imput observation(s) based in distribution.
- Correct observation(s) on physical grounds.
- Reject it (them) and possibly take additional observation(s).

Point Anomalies

Point Anomaly (Outlier)



Maastricht University

BISS Institute

Point Anomalies

Definition

Let the sample of n observations be denoted in order of increasing magnitude by $x_1 < x_2 \ldots < x_n$. Let x_n be the doubtful value, i.e. the largest value. The test criterion, T_n , recommended here for a single outlier is as follows:

$$T_n = \frac{(x_n - \bar{X})}{SD}$$

where
$\bar{x}$ = arithmetic average of all n values
and the estimator of SD is

$$SD = \sqrt{\frac{\sum_{i=o}^{n}(x_i - \bar{x})^2}{N-1}} = \sqrt{(\bar{X^2}) - (\bar{X})^2}$$

Point Anomalies

Correlation

$$Cor(X, Y) = \frac{Cov(X, Y)}{Std(X)Std(Y)}$$

The estimator is:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Standard Error (SE)

As well as the standard deviation (of the sampling distribution) of a statistic:

$$SE(X) = \frac{Std(X)}{\sqrt{n}}$$

Commonly considered for the mean with the estimator: $SE(\hat{x}) = \sigma_x/\sqrt{n}$

Point Anomalies

Feature Distribution

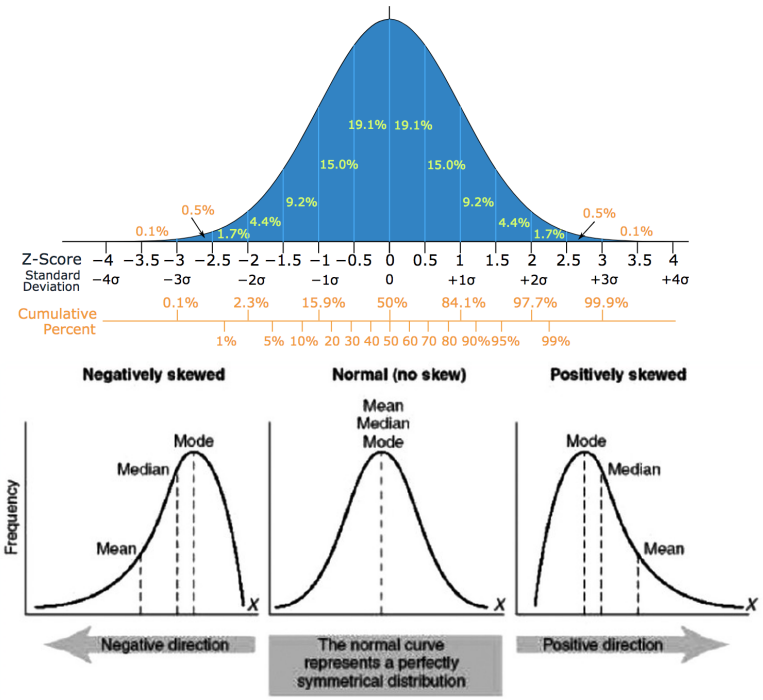
The law of large numbers (weak version) states that the sample average converges in probability towards the expected value.

$$\bar{X}_n \rightarrow \mu, \quad when \;\; n \rightarrow \text{inf}$$

- For any nonzero margin specified, no matter how small, with a sufficiently large sample, there will be a very high probability that the average of the observations will be close to the expected value.
- Applies on independent identically distributed random variables.
- If we have enough observations it is possible have an idea of the distribution of the variable plotting the density.

Point Anomalies

Density Plot



Collective Anomalies

Collective Anomalies

Definition

If a collection of related data instances is anomalous with respect to the entire data set, it is termed a collective anomaly. The individual data instances in a collective anomaly may not be anomalies by themselves, but their occurrence together as a collection is anomalous. The common technique to find this kind of anomalies is using clustering models.

Clustering

It refers to the task of grouping data so that points in the same cluster are **highly similar** to each other, while points in different clusters are dissimilar. Is a form of unsupervised learning because there is no target variable indicating which groups the training data belong to.

Collective Anomalies

Applications

- Social Network Analysis
- Intrusion Detection
- Process Mining
- Location Dependent Patterns

Note: Point anomalies can occur in any data set, collective anomalies can occur only in data sets in which data instances are related.

Collective Anomalies

Applications

Consider a sequence of actions occurring in a computer as shown below:

... http-web, buffer-overflow, http-web, http-web, smtp-mail, ftp, http-web, ssh, smtp-mail, http-web, **ssh, buffer-overflow, ftp**, http-web, ftp, smtp-mail,http-web ...

The highlighted sequence of events **(buffer-overflow, ssh, ftp)** correspond to a typical Web-based attack by a remote machine followed by copying of data from the host computer to a remote destination via ftp. It should be noted that this collection of events is an anomaly, but the individual events are not anomalies when they occur in other locations in the sequence.

k-Means Algorithmn

Is a method of vector quantization, and aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, K-means finds cluster centers by minimizing the sum of squared distances from each point to its assigned cluster. Points are assigned to the cluster whose center is closest.

After initial centers are chosen, two steps repeat until the cluster assignment no longer changes for any point (which is equivalent to the cluster centers no longer moving):

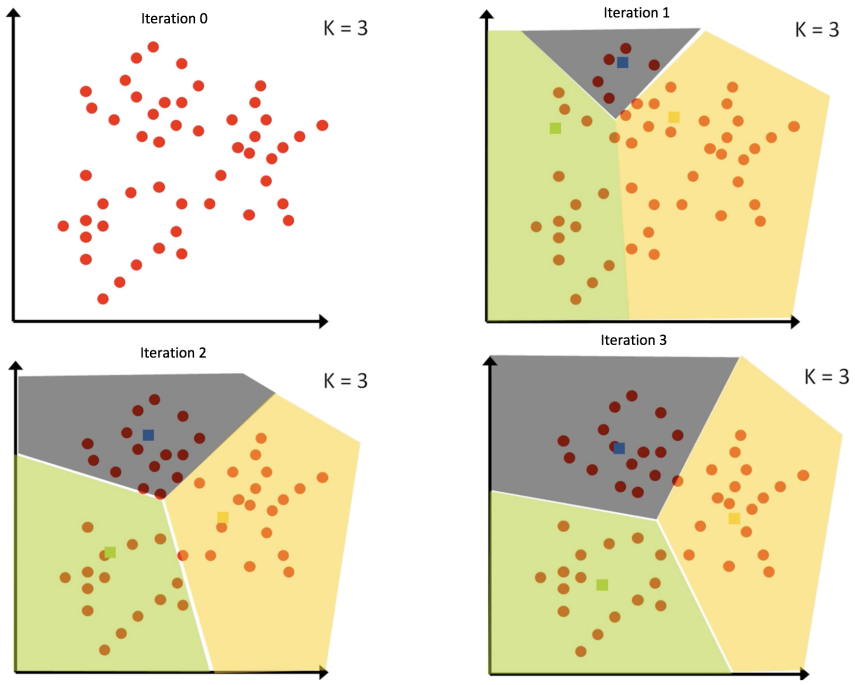
Assign each point to the cluster with the closest center. Update each cluster center to the be mean of the assigned points.

k-Means Algorithmn

- Input K set of points $x_1, \ldots, x_n$
- Place centroids $c_1, \ldots, c_k$ at random locations
- Repeat until converge:
  - for each pount x_i:
  - find nearest centroid c_j $argminD(x_i, c_j)$
  - assign the point x_i to cluster j
  - for each cluster $j = 1, \ldots, K$
  - new centroid c_j = mean of all points (coordinates) x_i assigned to the cluster in previous step

$$c_j(a) = \frac{1}{n} \sum_{x_i \to c_j} x_i(a)$$

Collective Anomalies



Collective Anomalies

Choosing k

In k-means for anomaly detection we must choose a value for k. This is still an active area of research and no definitive answers.

- The problem is much different than choosing a tuning parameter in regression or classification because there is no observable label to predict.
- The true risk R and estimated risk R_n decrease to 0 as k increases.
- Heuristic

$$K = minArg(\sqrt{\frac{N}{2}})$$

Collective Anomalies

Example

- Go to **k means crime** notebook
- Discuss data treatment
- Create a k means model for crime data
- Try for Airbnb data