KDnuggets

Subscribe to KDnuggets

Submit a blog
Win a Reward!

search KDnuggets Search

- Blog
- Opinions
- Tutorials
- Top stories
- Courses
- Datasets
- Education: Online
- Certificates
- Events / Meetings
- Jobs
- Software
- Webinars

Submit a blog to KDnuggets -- Top Blogs Win A Reward

Topics: AI | Data Science | Data Visualization | Deep Learning | Machine Learning | NLP | Python | R | Statistics

KDnuggets Home » News » 2017 » Apr » Tutorials, Overviews » Introduction to Anomaly Detection (17:n13)

Introduction to Anomaly Detection

<= Previous post
Next post =>

http likes 305
Like 0 Share 0 Tweet Share Share 56

Tags: Anomaly Detection, Datascience.com, Python, Time Series

This overview will cover several methods of detecting anomalies, as well as how to build a detector in Python using simple moving average (SMA) or low-pass filter.

By DataScience.com Sponsored Post.

This overview is intended for beginners in the fields of data science and machine learning. Almost no formal professional experience is needed to follow along, but the reader should have some basic knowledge of calculus (specifically integrals), the programming language Python, functional programming, and machine learning.

Introduction: Anomaly Detection

Anomaly detection is a technique used to identify unusual patterns that do not conform to expected behavior, called outliers. It has many applications in business, from intrusion detection (identifying strange patterns in network traffic that could signal a hack) to system health monitoring (spotting a malignant tumor in an MRI scan), and from fraud detection in credit card transactions to fault detection in operating environments.

This overview will cover several methods of detecting anomalies, as well as how to build a detector in Python using simple moving average (SMA) or low-pass filter.

Fig 1. Anomalies in Sunspots

What Are Anomalies?

Before getting started, it is important to establish some boundaries on the definition of an anomaly. Anomalies can be broadly categorized as:

1. **Point anomalies:** A single instance of data is anomalous if it's too far off from the rest. *Business use case:* Detecting credit card fraud based on "amount spent."
2. **Contextual anomalies:** The abnormality is context specific. This type of anomaly is common in time-series data. *Business use case:* Spending $100 on food every day during the holiday season is normal, but may be odd otherwise.
3. **Collective anomalies:** A set of data instances collectively helps in detecting anomalies. *Business use case:* Someone is trying to copy data form a remote machine to a local host unexpectedly, an anomaly that would be flagged as a potential cyber attack.

Anomaly detection is similar to - but not entirely the same as - noise removal and novelty detection. **Novelty detection** is concerned with identifying an unobserved pattern in new observations not included in training data - like a sudden interest in a new channel on YouTube during Christmas, for instance. **Noise removal** (NR) is the process of immunizing analysis from the occurrence of unwanted observations; in other words, removing noise from an otherwise meaningful signal.

Anomaly Detection Techniques

Simple Statistical Methods

The simplest approach to identifying irregularities in data is to flag the data points that deviate from common statistical properties of a distribution, including mean, median, mode, and quantiles. Let's say the definition of an anomalous data point is one that deviates by a certain standard deviation from the mean. Traversing mean over time-series data isn't exactly trivial, as it's not static. You would need a rolling window to compute the average across the data points. Technically, this is called a rolling average or a moving average, and it's intended to smooth short-term fluctuations and highlight long-term ones. Mathematically, an n-period simple moving average can also be defined as a "low pass filter." (A Kalman filter is a more sophisticated version of this metric; you can find a very intuitive explanation of it here.)

Challenges The low pass filter allows you to identify anomalies in simple use cases, but there are certain situations where this technique won't work. Here are a few:

- The data contains noise which might be similar to abnormal behavior, because the boundary between normal and abnormal behavior is often not precise.
- The definition of abnormal or normal may frequently change, as malicious adversaries constantly adapt themselves. Therefore, the threshold based on moving average may not always apply.
- The pattern is based on seasonality. This involves more sophisticated methods, such as decomposing the data into multiple trends in order to identify the change in seasonality.

Machine Learning-Based Approaches

Below is a brief overview of popular machine learning-based techniques for anomaly detection.

Density-Based Anomaly Detection

Density-based anomaly detection is based on the k-nearest neighbors algorithm.

Assumption: Normal data points occur around a dense neighborhood and abnormalities are far away.

The nearest set of data points are evaluated using a score, which could be Eucledian distance or a similar measure dependent on the type of the data (categorical or numerical). They could be broadly classified into two algorithms:

1. K-nearest neighbor: k-NN is a simple, non-parametric lazy learning technique used to classify data based on similarities in distance metrics such as Eucledian, Manhattan, Minkowski, or Hamming distance.
2. Relative density of data: This is better known as local outlier factor (LOF). This concept is based on a distance metric called reachability distance.

Clustering-Based Anomaly Detection

Clustering is one of the most popular concepts in the domain of unsupervised learning.

Assumption: Data points that are similar tend to belong to similar groups or clusters, as determined by their distance from local centroids.

K-means is a widely used clustering algorithm. It creates 'k' similar clusters of data points. Data instances that fall outside of these groups could potentially be marked as anomalies.

Support Vector Machine-Based Anomaly Detection

A support vector machine is another effective technique for detecting anomalies. A SVM is typically associated with supervised learning, but there are extensions (OneClassCVM, for instance) that can be used to identify anomalies as an unsupervised problems (in which training data are not labeled). The algorithm learns a soft boundary in order to cluster the normal data instances using the training set, and then, using the testing instance, it tunes itself to identify the abnormalities that fall outside the learned region.

Depending on the use case, the output of an anomaly detector could be numeric scalar values for filtering on domain-specific thresholds or textual labels (such as binary/multi labels).

Building a Simple Detection Solution Using a Low-Pass Filter

In this section, we will focus on building a simple anomaly-detection package using moving average to identify anomalies in the number of sunspots per month in a sample dataset, which can be downloaded here using the following command:

```
wget -c -b www-personal.umich.edu/~mejn/cp/data/sunspots.txt
```

The file has 3,143 rows, which contain information about sunspots collected between the years 1749-1984. Sunspots are defined as dark spots on the surface of the sun. The study of sunspots helps scientists understand the sun's properties over a period of time; in particular, its magnetic properties.

Read the rest of this tutorial including the Python code on **DataScience.com site**.

<= Previous post
Next post =>

Top Stories Past 30 Days

Most Popular

1. **Data Scientist, Data Engineer & Other Data Careers, Explained**
2. **Vaex: Pandas but 1000x faster**
3. **Data Science Books You Should Start Reading in 2021**
4. **Data Preparation in SQL, with Cheat Sheet!**
5. **Charticulator: Microsoft Research open-sourced a game-changing Data Visualization platform**

Most Shared

1. **A Guide On How To Become A Data Scientist (Step By Step Approach)**
2. **Data Scientist, Data Engineer & Other Data Careers, Explained**
3. **How to Determine if Your Machine Learning Model is Overtrained**
4. **DeepMind Wants to Reimagine One of the Most Important Algorithms in Machine Learning**
5. **Essential Linear Algebra for Data Science and Machine Learning**

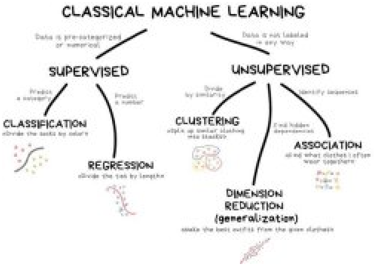
Latest News

- 5 Tasks To Automate With Python
- Beyond Brainless AI with a Feature Store
- 10 Deadly Sins of Machine Learning Model Training
- BigQuery vs Snowflake: A Comparison of Data Warehouse G...
- How a Data Scientist Should Communicate with Stakeholders
- Will There Be a Shortage of Data Science Jobs in the Ne...

Top Stories
Last Week

Most Popular

1. **A Guide On How To Become A Data Scientist (Step By Step Approach)**
2. **Top Programming Languages and Their Uses**
3. **Data Scientist, Data Engineer & Other Data Careers, Explained**
4. **Vaex: Pandas but 1000x faster**
5. **Choosing the Right BI Tool for Your Business**


CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
Find hidden dependencies
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

Most Shared

1. **A Guide On How To Become A Data Scientist (Step By Step Approach)**
2. **Top Programming Languages and Their Uses**
3. **How to Deal with Categorical Data for Machine Learning**
4. **Top Stories, May 17-23: Data Scientist, Data Engineer & Other Data Careers, Explained**
5. **Essential Machine Learning Algorithms: A Beginner's Guide**

More Recent Stories

- Will There Be a Shortage of Data Science Jobs in the Next 5 Ye...
- Similarity Search: Euclid of Alexandria goes shoe shopping
- Machine Learning Model Interpretation
- Stop (and Start) Hiring Data Scientists
- How to Make Python Code Run Incredibly Fast
- How to Create and Deploy a Simple Sentiment Analysis App via API
- How I Doubled My Income with Data Science and Machine Learning
- Overcoming the Simplicity Illusion with Data Migration
- Make Pandas 3 Times Faster with PyPolars
- Top 4 Data Extraction Tools

- Top Stories, May 24-30: A Guide On How To Become A Data Scient...
- Supercharge Your Machine Learning Experiments with PyCaret and...
- State of Mathematical Optimization Report, 2021
- Essential Math for Data Science: Basis and Change of Basis
- 4 Tips for Dataset Curation for NLP Projects
- Choosing the Right BI Tool for Your Business
- AIRSIDE LIVE Is Where Big Data, Data Security and Data Governa...
- Great New Resource for Natural Language Processing Research an...
- AI Books you should read in 2021
- Top Data and Analytics Trends

KDnuggets Home » News » 2017 » Apr » Tutorials, Overviews » Introduction to Anomaly Detection (17:n13)