ANOMALY DETECTION

Lesson 2: Extreme Value Analysis, Angle-based and Depth-based Techniques

Learning objectives

You will be able to:

- Describe probabilistic models for anomaly detection
- Apply extreme value analysis
- Apply angle-based and depth-based techniques
- Use Python* to perform anomaly detection on one- and two-dimensional data

Probabilistic models for anomaly detection

Introduction

- In Lesson 1, we introduced statistical tools for anomaly detection
- These were used heuristically (= rules of thumb)
- The statistical tools become more powerful if they can be connected to an underlying probabilistic model

You don't always have such a model, but if you do and believe it is reliable, use it.

Probabilistic models for anomaly detection

Workflow

- Choose an appropriate model for your data
- Select a probability threshold below which you will label the data an anomaly
- Calculate the probability of observing each instance your data
- Those instances which fall below the threshold are anomalies

Probabilistic models for anomaly detection

Example

- You are looking for match fixing in the 2018 World Cup soccer matches
- As a first test, you look at the number of goals in each match
- Call this number *n* (where n=0,1,2,3…)
- If the probability of observing *n* goals is below 2%, the match is an anomaly

Important: this threshold must be set BEFORE you look at the data.

Probabilistic models for anomaly detection

Example (continued)

- It has been shown that the number of goals in a World Cup match is well approximated by a Poisson distribution
- The probability of scoring *n* goals in a match is given by:

$$P(n) = \frac{\lambda^{n} e^{-\lambda}}{n!}$$

where λ is the average number of goals per match

Probabilistic models for anomaly detection

Example (continued)

- For World Cup events of the modern era, $\lambda = 2.5$
- Using the table of probabilities on the right, we see that matches with 7 or 8 goals would be labeled as anomalies [P(n) < 2%]
- In the 2018 World Cup, there were three matches with 7 goals and no matches with 8 goals:
  - Belgium 5-2 Tunisia
  - England 6-1 Panama
  - France 4-3 Argentina

n	P(*n*)
0	0.082
1	0.205
2	0.257
3	0.214
4	0.134
5	0.067
6	0.028
7	0.010
8	0.003

Probabilistic models for anomaly detection

Note of caution

- Finding anomalies DOES NOT mean you detected match fixing
- It means you should look at the anomalous matches more closely
- It is still possible that the matches with 7 goals happened by chance
- Other sources of problems with probabilistic models for anomaly detection:
  - model is inappropriate
  - parameters are wrong
  - test statistic is poorly chosen

EXTREME VALUE ANALYSIS

Extreme value analysis

Motivation

- Sometimes an anomaly is an extreme event: a very big insurance loss, a very large flood, a very hot summer, etc.
- As such events can be catastrophic, it is natural to ask how likely are these extreme events
- Problem: extreme events are rare, so are hard to model with typical probability distribution because there is very little data

Extreme value analysis

The challenge with modeling extreme events

- From the World Cup matches discussed previously, we can calculate the probability of scoring 12 goals in a match: it is about 1 in 100,000
- Sounds very, very unlikely, but it happened
- The probability estimate is suspect. Why? The data used to calculate the parameter λ only includes matches with 0 to 8 goals scored. It is unlikely to capture the behavior of extreme events

How do you predict events outside the range of observations?

Extreme value analysis

Two practical approaches*

- Block maxima: take one maximum value per unit time (often annual)
- Peaks over threshold (exceedances): take all values over a specific threshold
- These approaches work because both the maxima and the exceedances are described by specific families of probability distributions

*Note: for simplicity, we will assume the extreme event is a maximum, but the approach also works when the extreme event is a minimum

Univariate extreme value analysis

One dimensional case

- To show how these approaches work, we will start with one-dimensional data
- Consider a sequence of independent and identically-distributed random variables

$$X_1, X_2 \ldots X_n$$

- Example: X_i is the daily ozone level on day i (we will work with this example in the Python* notebook that accompanies this lecture)

Generalized extreme value distribution

The extreme value theorem

- Let the maxima of the random variables be $M_n = \max(X_1, X_2 \ldots X_n)$
- When *n* is large, the distribution of the maxima M_n is a generalized extreme value (GEV) distribution, which is characterized by three parameters:

$$x = \text{shape (type)}$$

$$m = \text{location}$$

$$s = \text{scale } (> 0)$$

GEV distribution: probability density function

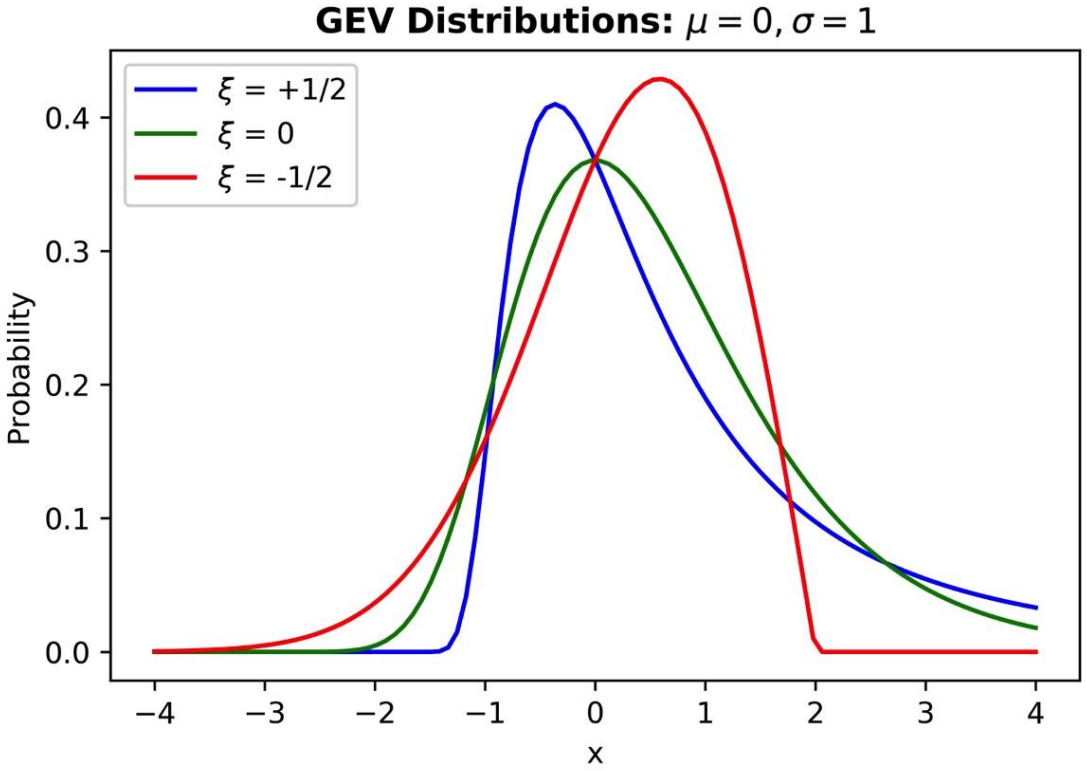
$$P(M_n = x) = \frac{1}{s} s(x)^{x+1} e^{-s(x)}$$

$$s(x) = \begin{cases} \left[1 + x\left(\frac{x - m}{s}\right)\right]^{-1/x} & \text{for } x \neq 0 \\ \exp\left[-\left(\frac{x - m}{s}\right)\right] & \text{for } x = 0 \end{cases}$$

$$x \in [m - \frac{s}{x}, +\infty) \text{ for } x > 0$$

$$x \in (-\infty, +\infty) \text{ for } x = 0$$

$$x \in (-\infty, m - \frac{s}{x}] \text{ for } x < 0$$


GEV Distributions: $\mu = 0, \sigma = 1$
ξ = +1/2
ξ = 0
ξ = -1/2
Probability
0.0
0.1
0.2
0.3
0.4
−4
−3
−2
−1
0
1
2
3
4
x

Block maxima

What does the extreme value theorem mean?

- As long as the underlying distribution of your data is not too strange, then regardless of what this distribution is, maxima of samples of size *n* will follow a GEV distribution if *n* is large enough
- So if you have enough data, you can use it to determine the three parameters that describe your GEV (ξ, μ, σ) = (shape, location, scale)
- Once you have your complete GEV (with parameters), you can answer questions such as, “How likely is it to exceed a certain value in a given unit of time?”

Block maxima

Workflow

- Divide your data into blocks of fixed size. Typically, the size is one year
- For each one year block, find the maximum value. For a yearly division, the collection of maxima is known as an “annual maximum series” (AMS)
- Fit the AMS data with a GEV distribution. Extract the shape, location, scale and parameters

Where judgment is needed: how to divide the data into blocks

(We will explore this point in the Python* notebook that accompanies this lecture)

Peaks over threshold

When you are interested in more than just a maximum value

- For example, consider ozone levels mentioned previously. An air quality index (AQI) for ozone over 200 is considered ‘very unhealthy’
- This level can be exceeded several times per year and there can be several years when it isn’t exceeded
- In this case, the block maxima approach isn’t useful
- Instead, you want the the probability of exceeding some threshold. This can be obtained with the Peak Over Threshold (POT) approach

Generalized Pareto Distribution

Pickands-Balkema-de Haans theorem

- Consider a sequence of independent and identically-distributed random variables
- Take only observations that are above a fixed threshold *u*
- When *u* is very large, the distribution of values above the threshold (exceedances) is a generalized Pareto distribution (GPD), which is characterized by three parameters:

$$x = \text{shape}\ ,\ m = \text{location},\ s = \text{scale}\ (>0)$$

GPD distribution: probability density function

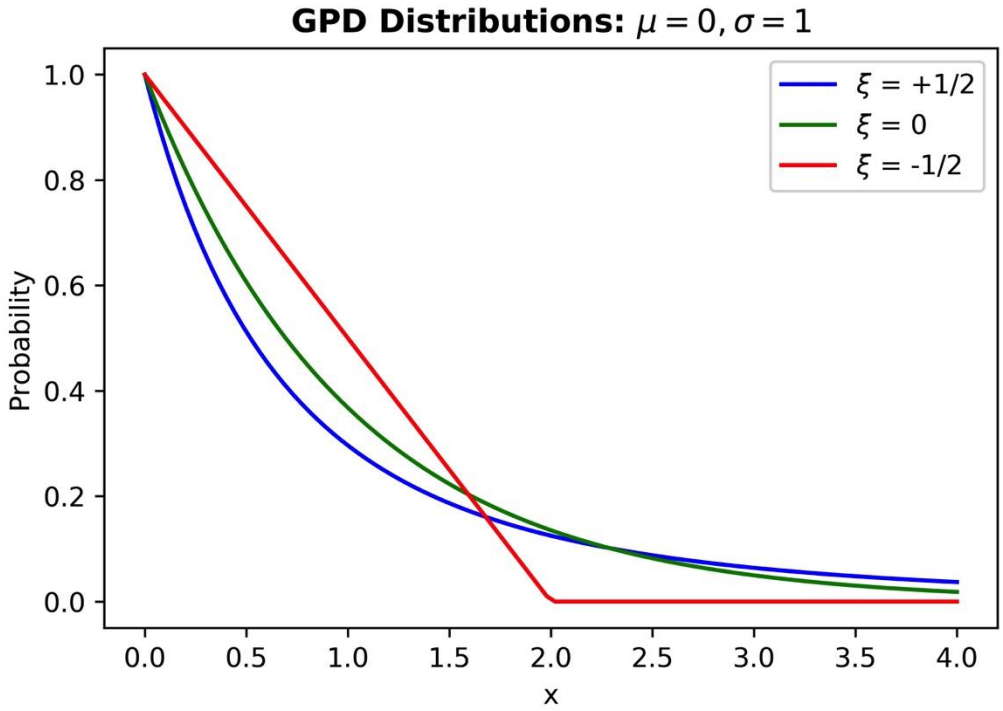
$$f(x) = t(x) / s$$

$$t(x) = \begin{cases} \left[1 + x\left(\frac{x - m}{s}\right)\right]^{\left(-\frac{1}{x} - 1\right)} & \text{if } x \neq 0 \\ \exp\left[-\left(\frac{x - m}{s}\right)\right] & \text{if } x = 0 \end{cases}$$

$x \in [m, +\infty)$ for $x > 0$

$x \in [m, m - \frac{s}{x}]$ for $x < 0$

$x \in [m, +\infty)$ for $x = 0$


GPD Distributions: $\mu = 0, \sigma = 1$
ξ = +1/2
ξ = 0
ξ = -1/2
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
x

Peak over threshold

What does the Pickands-Balkema-de Haans theorem mean?

- Universality—almost all probability distributions have a tail that is a GPD
- So if you have enough data, you can use it to determine the three parameters that describe your GPD (ξ, μ, σ) = (shape, location, scale)
- Note that the theorem holds in the limit of an infinite threshold *u*
- In practice, you must choose a finite *u* and there is a tradeoff:
  - *large u, theorem applies better, but have few data points (poor statistics)*
  - *small u, have more data points, but theorem is less applicable*

Peaks over threshold

Workflow

- Choose a threshold
- Use only the data that is above threshold
- Fit this data with a GPD distribution. Extract the shape, location, scale and parameters

Where judgment is needed: how to choose the threshold

(We will explore this point in the Python* notebook that accompanies this lecture)

Multivariate extreme value analysis

Generalization of univariate case

- Can generalize both block maxima and peaks over threshold approaches to higher dimensions
- The univariate GEV distributions and GDP (in appropriate combinations) are still useful
- New twist: correlation between the variables
- Numerous models exist to capture correlations, all of which introduce additional fitting parameters

Extreme value analysis (EVA): final thoughts

Two notes of caution

- For EVA to give reasonable results, it is important to check that mathematical assumptions behind the theorems are met—e.g., time series must be stationary
- EVA will detect anomalies if they are maxima or minima. But what if the anomalies aren't either?

Anomaly is a maximum

Anomaly is neither maximum nor minimum

Anomaly detection without probability distributions

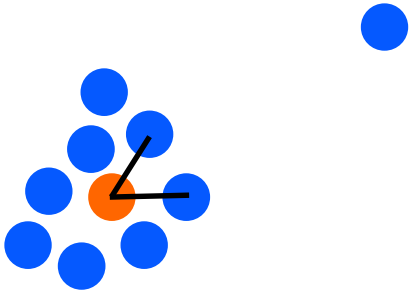
The geometry of the data

- Until now we focused on ways to detect anomalies using probability distributions
- This approach has many strengths, but isn't always applicable. Perhaps you don't know the underlying probabilistic model or the data doesn't satisfy the assumptions of extreme value analysis
- In such cases, it is useful to consider techniques based on the geometry (spatial structure) of the data
- Here we will discuss two such techniques for multivariate data: angle-based and depth-based

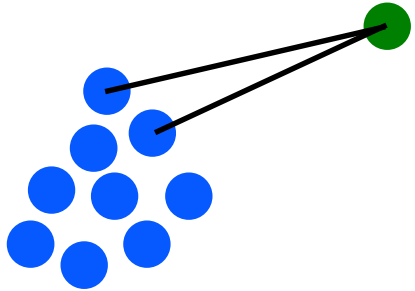
ANGLE-BASED TECHNIQUES

Angle-based techniques

The essential idea

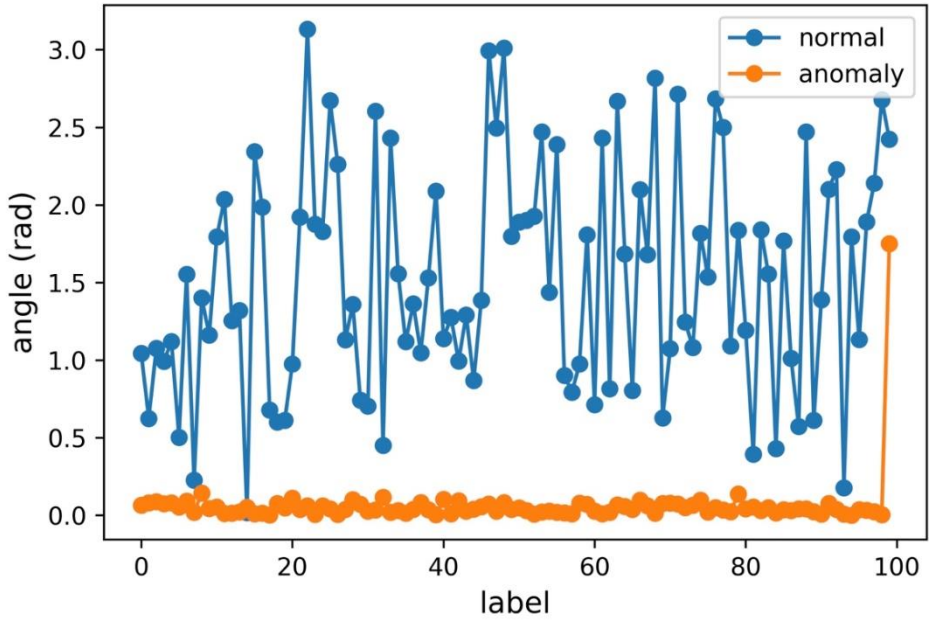


For a normal point (orange), the angle it makes with any other two data points varies a lot as you choose different data points



For an anomaly (green), the angle it makes with any other two data points doesn't vary much as you choose different data points

Angle-based techniques



Angle-based techniques

Implementation

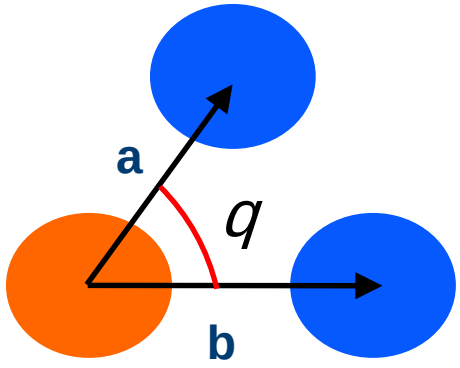
- For each data point, determine the angle it makes with all pairs of other data points
- Calculate the variance of this angle
- Points for which the variance is below are predetermined threshold are anomalies

As described, this algorithm is very slow for large datasets

(We will discuss the runtime complexity of the algorithm in the Python* notebook that accompanies this lecture)

Calculating the angle

Use vector distances between points



Dot product

$$\cos q = \frac{\boldsymbol{a} \times \boldsymbol{b}}{ab}$$

Magnitude of **a** Magnitude of **b**

Other angle-like metrics

Improving the performance of angle-based techniques

- The angle is a good metric in principle, it doesn't always work well in practice
- Other angle-like metrics have been devised. For example: $\cos\psi = \frac{\boldsymbol{a} \times \boldsymbol{b}}{a^2 b^2}$
  - While ψ is often referred to as angle, it isn't a true angle
  - Note the square powers in the denominator, which introduce a distance-dependence in this metric

(We will examine these points in the Python* notebook that accompanies this lecture)

DEPTH-BASED TECHNIQUES

Depth-based techniques

Another approach that doesn't use a probability distribution

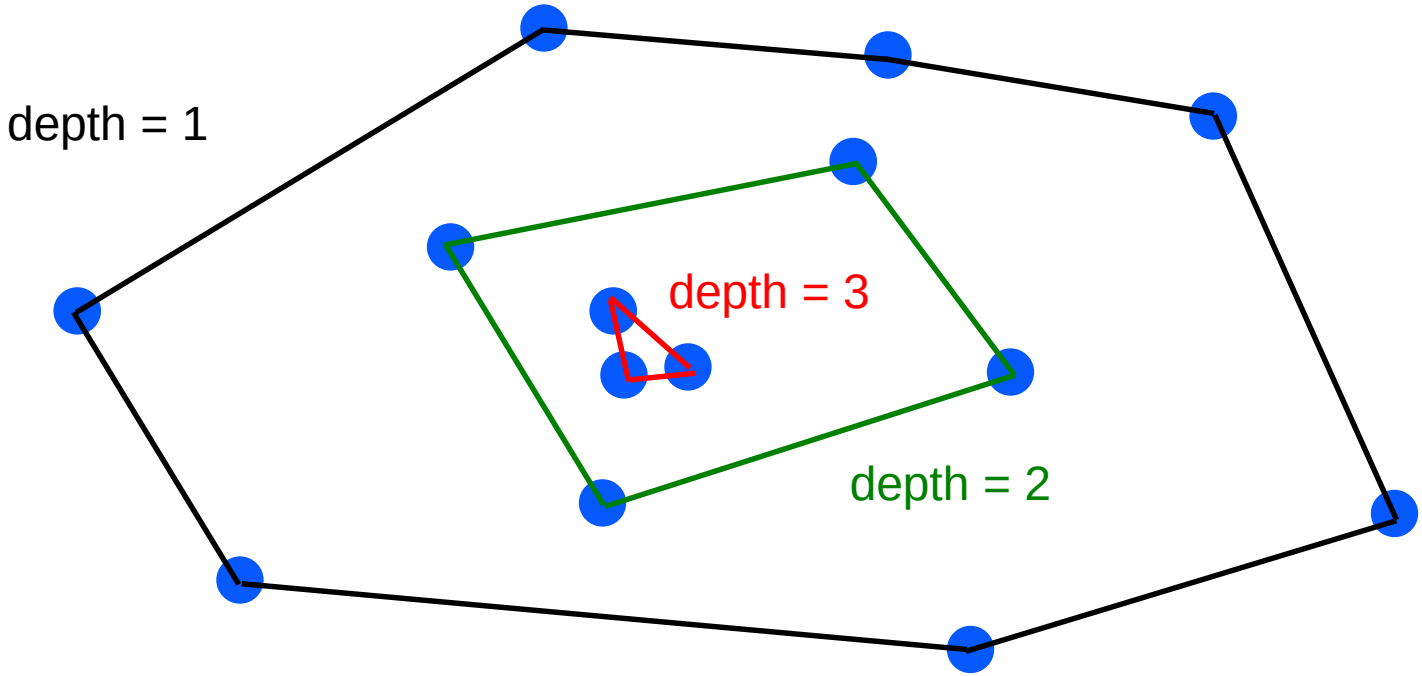
- Anomalies are assumed to lie at the edge of the data space
- Organize the data in layers
- Each layer in labeled by its depth. The outermost layer is depth = 1, the next is depth = 2 and so on
- Anomalies are those points with a depth below a pre-determined threshold

Convex hull

A common depth-based approach

- Convex hull: the smallest convex set that contains the data
- Points on the convex hull of the whole data space have depth = 1
- Points on the convex hull of the dataset after removing all of the depth = 1 points have depth = 2
- And so on…
- Anomalies are points with a depth ≤ *n* (where *n* is a positive integer)
- Natural implementation as a recursive algorithm (see Python* notebook)

Convex hull



Convex hull

Comments

- Convex hull is typically efficient only for two- and three-dimensional data
- While the algorithm is usually used to classify data (anomaly vs. normal), the depth can also be used as a scoring mechanism
- Not suitable for anomaly detection if the anomalies aren't at edges of data

CONCLUSION

Use Python* for anomaly detection

Next up is a look at applying these concepts in Python*

- See notebook entitled *Extreme_Anomaly_Detection_student.ipynb*

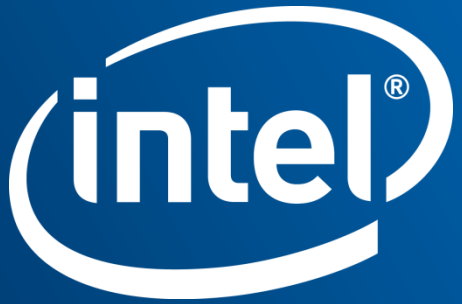
Learning objectives recap

In this session you learned how to:

- Describe probabilistic models for anomaly detection
- Apply extreme value analysis
- Apply angle-based and depth-based techniques
- Use Python* to perform anomaly detection on one- and two-dimensional data

References

- *An Introduction to Statistical Modeling of Extreme Values* by S. Coles (Springer-Verlag 2001)
- *Angle-Based Outlier Detection in High-Dimensional Data* by H.-P. Kriegel, M. Schibert, A. Zimek (2008)
- *Outlier Detection Techniques* by H.-P. Kriegel, P. Kröger, A. Zimek (2010)


intel®