experience what's inside™

ANOMALY DETECTION

Lesson 1: Introduction

Learning objectives

You will be able to:

- Define various types of anomalies
- Discuss the applications of anomaly detection
- Explain the basic statistics related to anomaly detection
- Use Python* to apply anomaly detection to one-dimensional data

What is an anomaly?

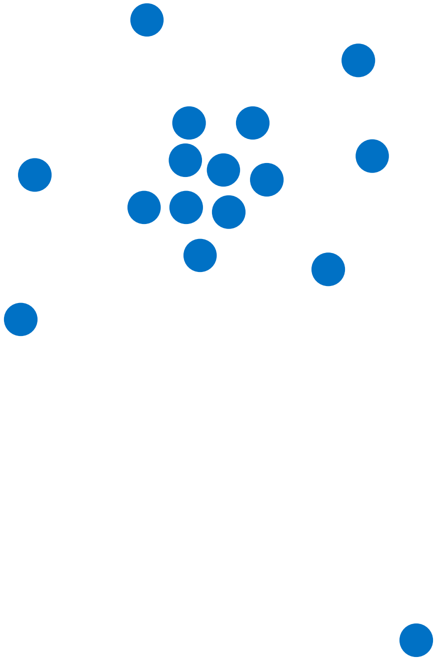
Data that differs a lot from the rest.

- An anomaly is “an observation which deviates so much from other observations as to arouse suspicions that it was generated by a different mechanism.” (Hawkins 1980)
- Also called “abnormality” or “deviant.”
- “Outlier” is also used as a synonym, but here we will use a more precise definition.

What is an anomaly (continued)?

Anomalies are a subset of outliers (Aggarwal 2013)

- All observations = normal data + outliers
- Outliers = noise + anomalies
- Noise = uninteresting outliers
- Anomaly = sufficiently interesting outlier



Anomalies: two fundamental questions

How big must the deviation be for a point to be classified as an anomaly?

- No easy answer. The classification depends in part on subjective judgment.

How do I separate an anomaly from noise?

- Depends on what is “sufficiently interesting” for you.

Types of anomalies

- **Point anomalies**: an individual data point seems strange when compared with the rest of the data. *Example*: an unusually large credit card purchase
- **Contextual anomalies**: the data seems strange in a specific context, but not otherwise. *Example*: a US credit card holder makes a purchase in Japan
- **Collective anomalies**: a collection of data points seems strange when compared with entire dataset, although each point may be OK. *Example*: ten consecutive credit card purchases for a sandwich at hourly intervals

Applications and use cases

- Fraud detection in credit card purchases
- Intrusion detection in computer networks
- Fault detection in mechanical equipment
- Earthquake warning
- Automated surveillance
- Monitoring gene expression for cancer classification
- Detect fake social media accounts

Anomaly detection: the fundamental idea

The approach used by almost all anomaly detection algorithms

- Create a model for what normal data should look like*
- Calculate a score for each data point that measures how far from normal it is
- If score is above a previously specified threshold, classify point as an anomaly

Devising an appropriate model and score is essential

*Note: “normal” is used here in the sense of “typical” or “usual,” which may or may not be related to the normal distribution.

Anomaly detection: modeling the data

The approach you take depends on what you know

- If you have examples of normal or anomalous data, you can use this information to find anomalies
  - Supervised anomaly detection (lesson 6)
- If you don't have any prior information about normal or anomalous data, you have to use a different approach
  - Unsupervised anomaly detection (this lesson and several others)
  - Requires probability and statistics to look for anomalies

REVIEW OF PROBABILITY AND STATISTICS

Probability distribution

- The chance of obtaining a data value (or range of values)
- The normal (Gaussian) distribution is the probability distribution most commonly used to model data
- Caution: while it is mathematically convenient and easy to use, the normal distribution may not be appropriate for your specific data. Do NOT use it without thinking about your data first.

Cumulative distribution function (CDF)

- For a real-valued random variable *X*, the CDF evaluated at *x* is the probability that *X* will take a value less than or equal to *x*
- Usually denoted as *F*(*x*). Four basic properties:

$$0 \le F(x) \le 1 \ \text{ for all } x$$

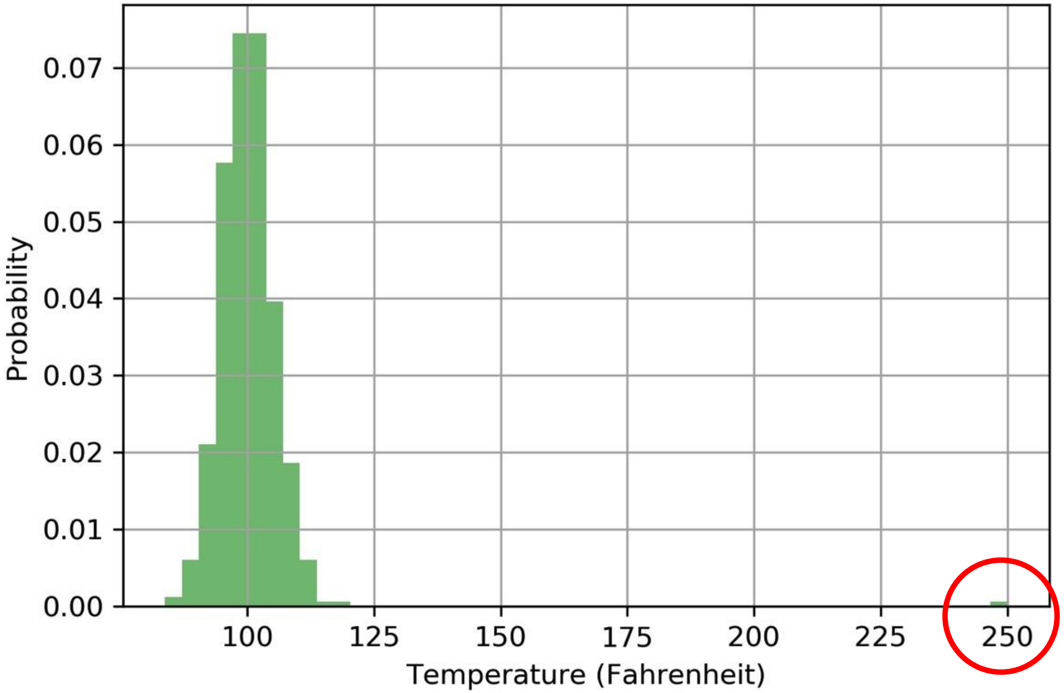
$$\lim_{x \to -\infty} F(x) = 0$$

$$\lim_{x \to +\infty} F(x) = 1$$

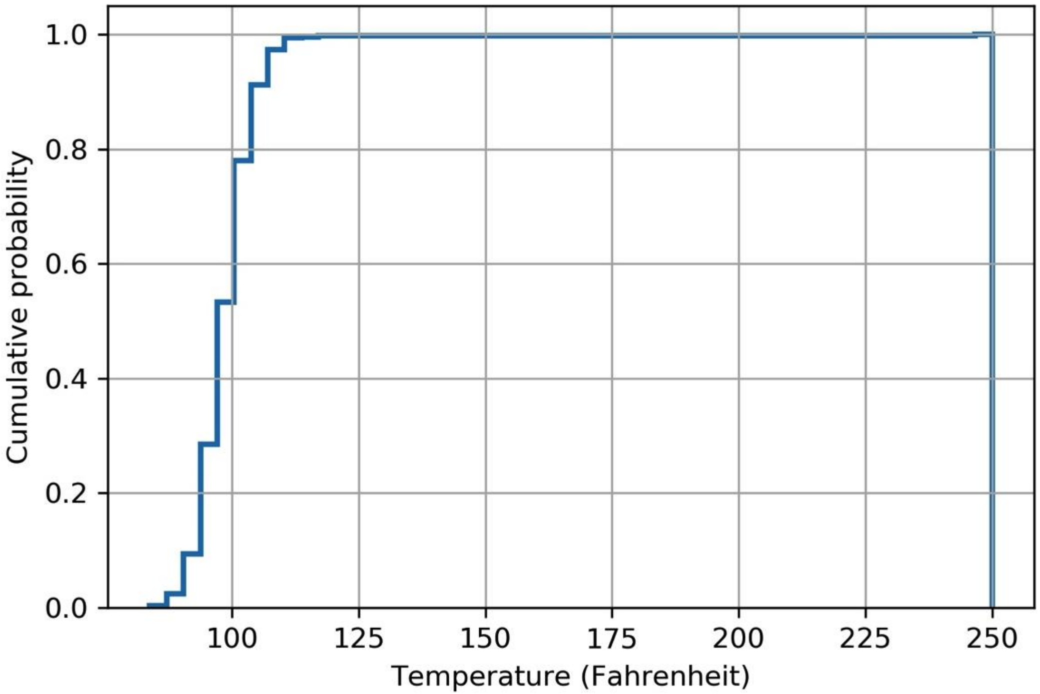
$$F(x) \text{ is a non-decreasing function of } x$$

Probability distribution vs. CDF

- Both are useful for anomaly detection
- If you want to identify anomalies as low probability events, then using a probability distribution is straightforward
- For visual inspection of anomalies, the CDF is often more robust



Can you see the anomaly?
If so, what is its value?



Outlier is at 250 °F

Fundamental statistics: mean

A type of average

- Mean: also known as the expected value
- For a discrete random variable X that can assume values $x_1, x_2, \ldots x_n$, it is given by

$$\mu = \mathrm{E}[X] = \sum_{i=1}^{n} x_i p(x_i)$$

- Here $p(x_i)$ is the probability of getting outcome x_i where $i = 1, 2, \ldots n$

Fundamental statistics: median and mode

Other types of averages

- Median: the value separating the higher half and lower half of the data
- Mode: the value that appears most often

Median and mode are usually less affected by outliers than mean

Fundamental statistics: example

Assume values all have equal probability

- values = *2, 2, 3, 4, 7, 8, 9*
  - *mean = 5; median = 4; mode = 2*
- Now introduce an outlier:
- values = *2, 2, 3, 4, 7, 8, 30*
  - *mean = 8; median = 4; mode = 2*

Mean changes, but median and mode do not

Fundamental statistics (continued)

The spread of the data about the mean

- Variance: the expected value of the square of the deviation of a random variable from its mean
  - For a discrete random variable:

$$\mathrm{var}(X) = \sigma^2 = \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] = \sum_{i=1}^{n} (x_i - \mu)^2 \, p(x_i)$$

Here σ is the standard deviation. It is used frequently in anomaly detection.

The value of σ is sensitive to the presence of anomalies

Fundamental statistics (continued)

Multivariate data

- Covariance: it measures the joint variability of two random variables

$$\operatorname{cov}(X,Y) = \mathrm{E}\left[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])\right] = \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y]$$

The covariance of a variable with itself is just the variance

$$\operatorname{cov}(X,X) = \operatorname{var}(X)$$

Fundamental statistics (continued)

Multivariate data

- Consider a vector of random variables:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

- Can construct a covariance matrix Σ whose entries are given by:

$$\Sigma_{ij} = \mathrm{cov}(X_i, X_j)$$

- The covariance matrix represents the generalization of variance to higher dimensions. It is often used in anomaly detection.

Statistical tests

Scoring anomalies

- A common method for scoring anomalies in 1D data is the z-score
- If the mean and standard deviation are known, then for each data point calculate the z-score as

$$z_i = \frac{x_i - m}{s}$$

- The z-score measures how far a point is away from the mean as signed multiple of the standard deviation
- Large absolute values of the z-score suggest an anomaly

Statistical tests

A note of caution

- Since the mean and standard deviation are themselves sensitive to anomalies, the z-score can sometimes be unreliable
- The modified z-score tackles this problem by using medians instead:

$$y_i = \frac{x_i - \tilde{X}}{\text{MAD}}$$

$$\tilde{X} = \text{median of } X$$

$$\text{MAD} = \text{median}(|x_i - \tilde{X}|)$$

- MAD = **m**edian **a**bsolute **d**eviation from the median
- Large absolute values of the modified z-score suggest an anomaly

Example: normal data

Dataset 1	z-score	mod z-score
2	-1.1	-1
2	-1.1	-1
3	-0.7	-0.5
4	-0.4	0
7	0.7	1.5
8	1.1	2
9	1.5	2.5

Mean	5
Std deviation	2.7
Median	4
MAD	2

Example: data with an anomaly

Dataset 2	z-score	mod z-score
2	-0.6	-1
2	-0.6	-1
3	-0.5	-0.5
4	-0.4	0
7	-0.1	1.5
8	0	2
30	2.4	13

Mean	8
Std deviation	9.2
Median	4
MAD	2

Statistical tests

Multivariate data

- The higher dimensional analog of the z-score is the Mahalanobis distance
- The Mahalanobis distance *d* of a data point from a set of observations is given by

$$d(\boldsymbol{X}) = \sqrt{(\boldsymbol{X} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{X} - \boldsymbol{\mu})}$$

$\boldsymbol{X}$ is the data point (a column vector)

$\boldsymbol{\mu}$ is the column vector of means

Σ is the covariance matrix

- Commonly used for anomaly detection
- Requires that the inverse covariance matrix exist (can be a problem)
- More robust versions of this distance have been devised

Use Python* for anomaly detection

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

- See notebook entitled *Introduction_to_Anomaly_Detection_student.ipynb*

Learning objectives recap

In this session you learned how to:

- Define various types of anomaly
- Discuss the applications of anomaly detection
- Explain the basic statistics related to anomaly detection
- Use Python* to apply anomaly detection to one-dimensional data

References

- *Identification of Outliers* by D.M. Hawkins (Champan & Hall 1980)
- *Outlier Analysis* by C.C. Aggarwal (Springer 2013)
  - First chapter available free
- *Eureka Statistics*