Anomaly Detection
Introduction - basics of anomaly detection

Paul Irofti
Cristian Rusu
Andrei Pătraşcu

Computer Science Department
University of Bucharest

Class structure

Topics for today:

- give a definition of anomaly detection
- provide some characteristics of anomaly detection
- analyze a simple example: anomaly detection with z-scores

A definition

What is an anomaly?

A definition

An outlier is an observation which deviate so much from the other observations as to arouse suspicions that it was generated by a different mechanism.[1]

[1] D. Hawkins. Identification of Outliers, Chapman and Hall, 1980.

What is an anomaly

The following are used interchangeably in the literature:

- anomaly detection
- outlier detection
- novelty detection
- intrusion detection

It is an ill-posed problem.

Hard to give a precise definition, hard to distinguish sometimes from noise

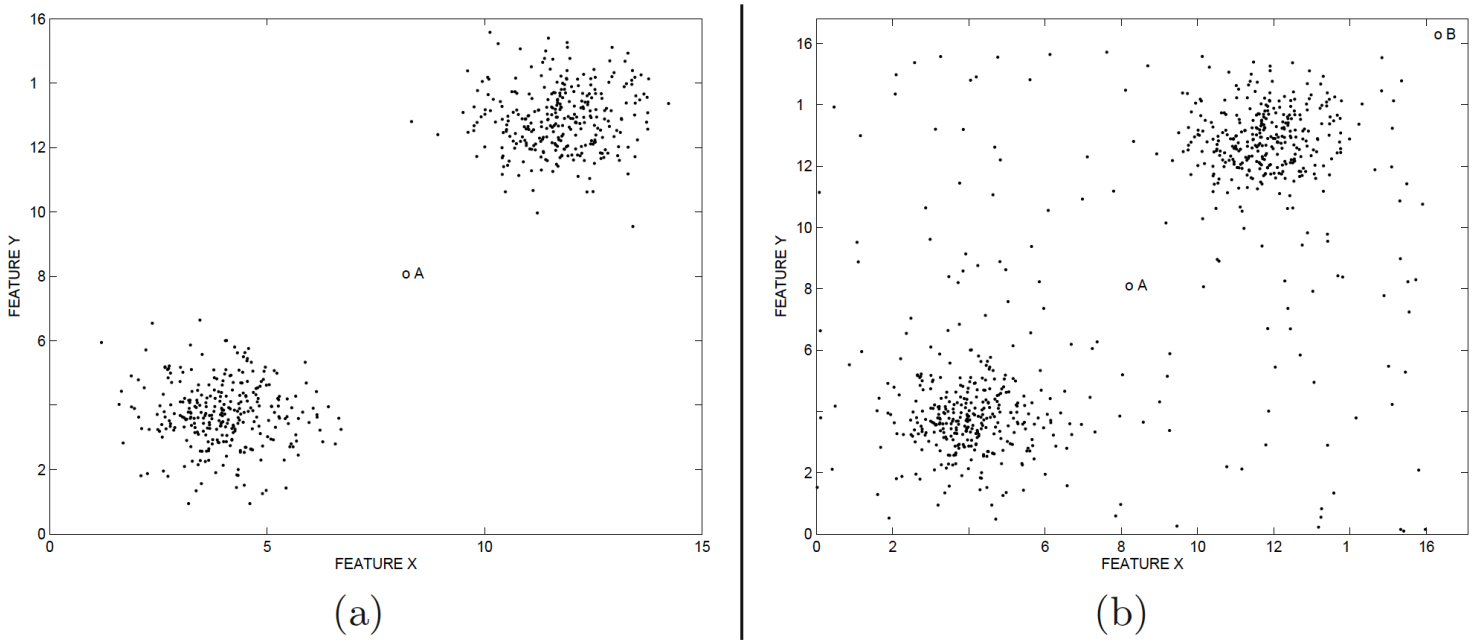

Figure: Figure 1.1 from Aggarwal 2013. On the right, the anomaly is hidden in the noise.

The model

One of the key ideas in anomaly detection is to create a model of the data and then find points in the dataset that are *far away* from the model

Do you see any problems with this idea?

The model

One of the key ideas in anomaly detection is to create a model of the data and then find points in the dataset that are *far away* from the model

Classic machine learning trade-off:

- if the model is too simple, then everyone is an anomaly;
- if the model is too large (over-parametrized) then you start to fit anomalies and noise

The model

One of the key ideas in anomaly detection is to create a model of the data and then find points in the dataset that are *far away* from the model

Machine learning models:

- Probabilistic and Statistical Models (EM approach)
- Linear Models
- Spectral Models
- Information Theoretic Models
- Meta Models

A statistical model

Assume that your data points are drawn from a Gaussian distribution

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

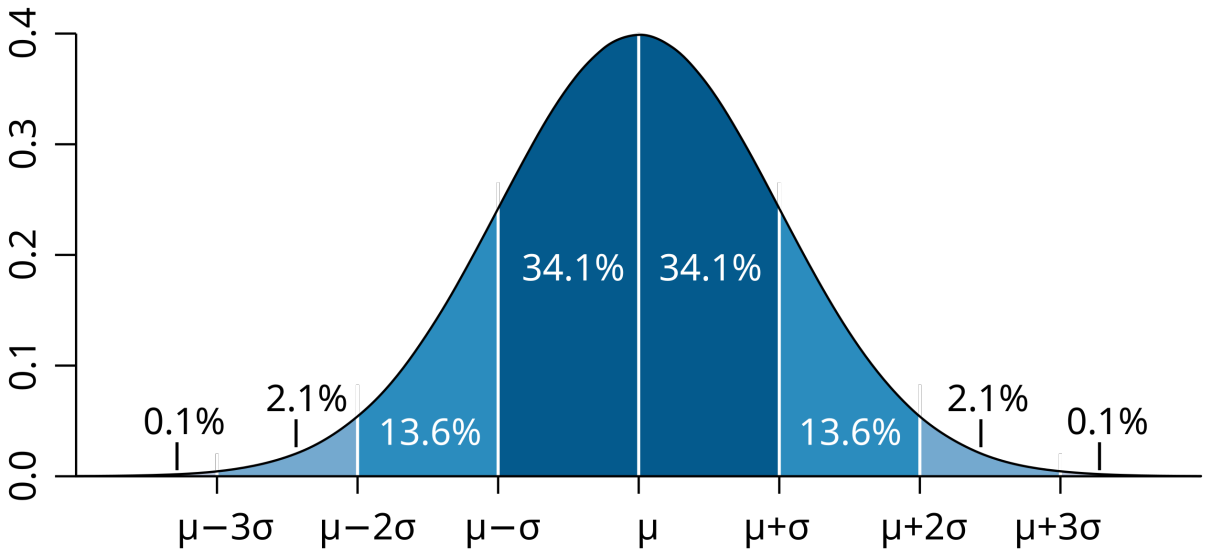

Figure: 1D Gaussian, source: wikipedia.

How do we define an anomaly in this case?

A statistical model

Assume that your data points are drawn from a Gaussian distribution

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

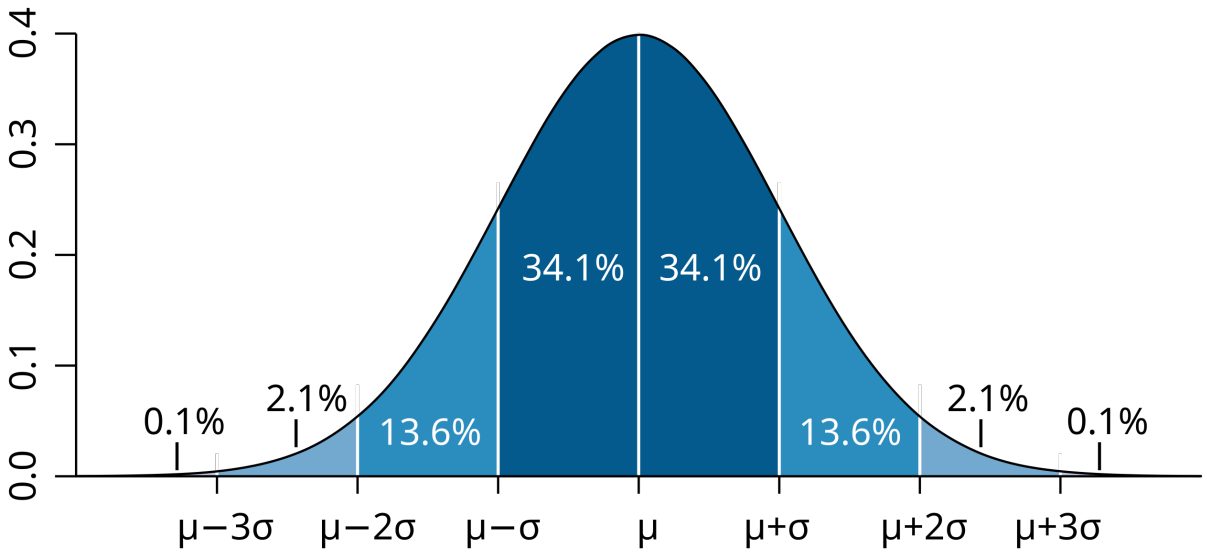

Figure: 1D Gaussian, source: wikipedia.

How do we define an anomaly in this case? z-score$(x, \mu, \sigma) = \frac{|x-\mu|}{\sigma}$

A statistical model

Assume that your data points are drawn from a Gaussian distribution

$$p(\mathbf{x}, \mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\}$$

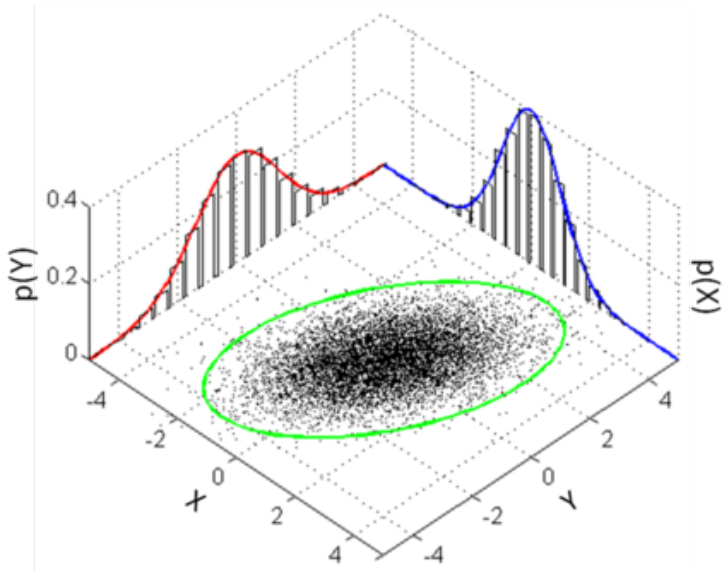

Figure: 2-dimnensional Gaussian, source: wikipedia.

How do we define an anomaly in this case?

A statistical model

Assume that your data points are drawn from a Gaussian distribution

$$p(\mathbf{x}, \mu, \Sigma) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\}$$

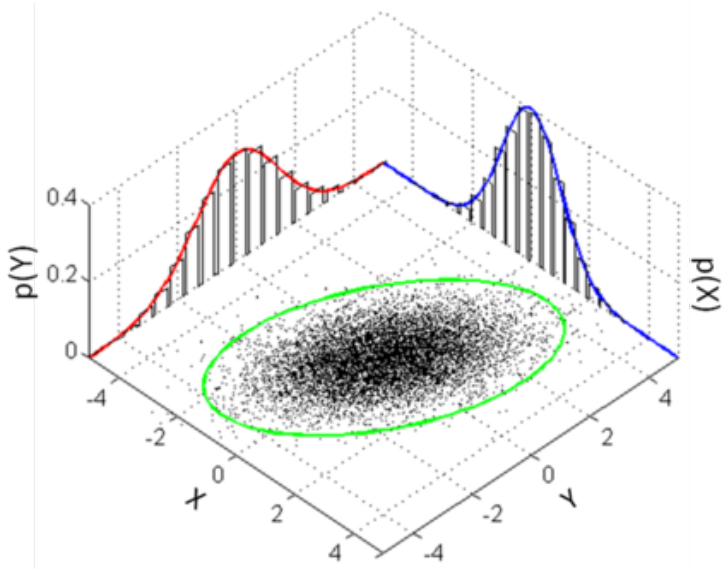

Figure: 2-dimnensional Gaussian, source: wikipedia.

How do we define an anomaly in this case?
z-score$(\mathbf{x}, \mu, \Sigma) = \sqrt{(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$

The end.