ALGORITHMS FOR ANOMALY DETECTION

Michael Davis, Ph.D.

CERN, Meyrin

6 & 7 March 2017

Artefacts

"I didn't even realize you could HAVE a data set made up entirely of outliers."

Introduction

One of the first topics of interest in study of statistics

> Astronomers prefer to reject completely observations which they judge to be too wide of the truth, while retaining the rest… I see no way of drawing a dividing line between those that are to be utterly rejected and those that are to be wholly retained; it may even happen that the rejected observation is the one that would have supplied the best correction to the others… I think each and every observation should be admitted whatever its quality…
>
> — Daniel Bernouilli
> *The most probable choice between several discrepant observations and the formation therefrom of the most likely induction* (1777)

Hadlum vs. Hadlum (1949)

- In August 1944, Mr. Hadlum left home on a period of military service
- 349 days later (12 Aug 1945), Mrs. Hadlum gave birth
- On his return, Mr. Hadlum filed for divorce

Hadlum vs. Hadlum (1949)

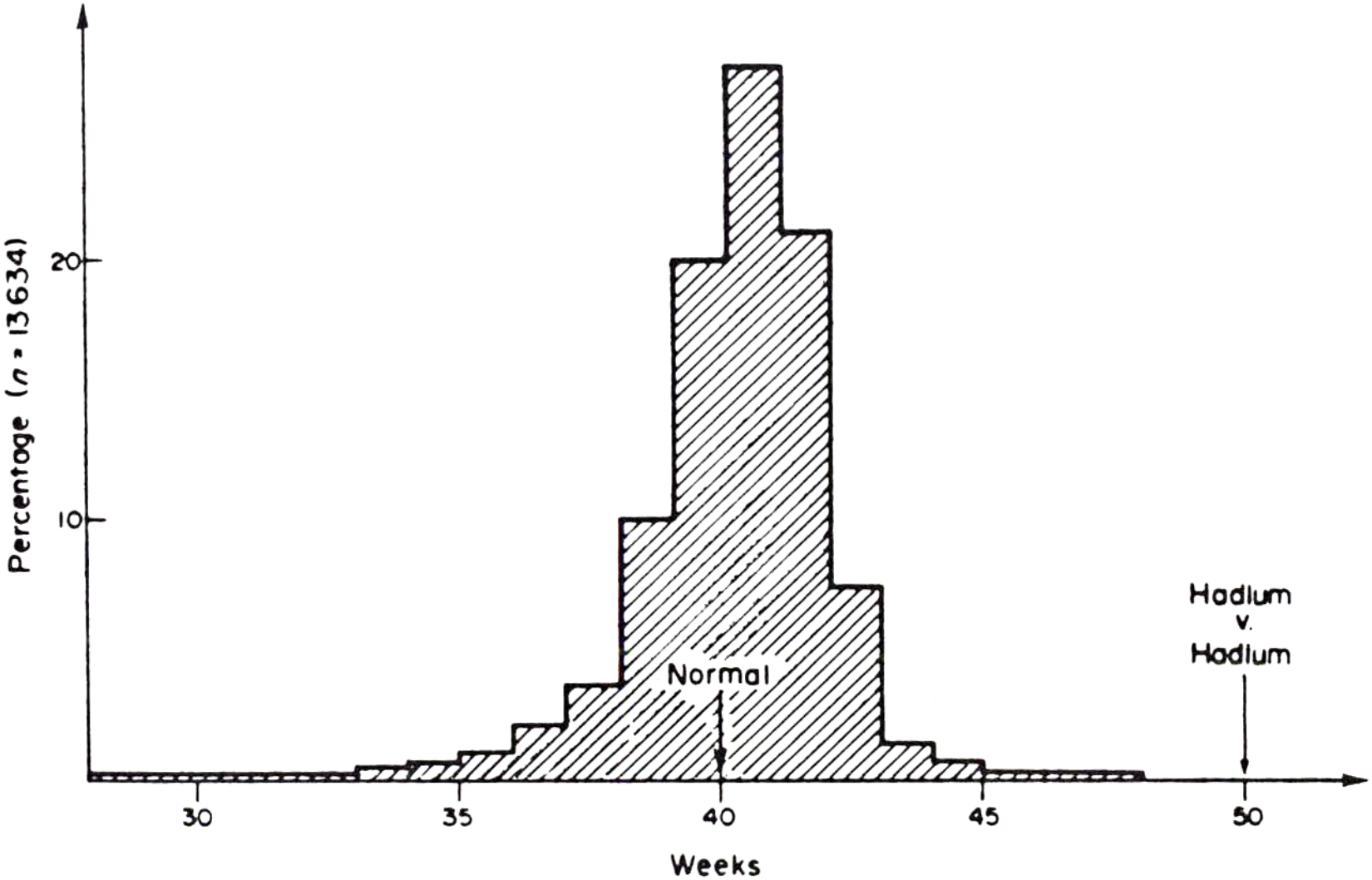
- In August 1944, Mr. Hadlum left home on a period of military service
- 349 days later (12 Aug 1945), Mrs. Hadlum gave birth
- On his return, Mr. Hadlum filed for divorce

Hadlum vs. Hadlum (1949)

- In August 1944, Mr. Hadlum left home on a period of military service
- 349 days later (12 Aug 1945), Mrs. Hadlum gave birth
- On his return, Mr. Hadlum filed for divorce

Hadlum vs. Hadlum (1949)

Histogram of Human Gestation Periods



Hadlum vs. Hadlum (1949)

- Average human gestation period is 280 days
- Mr. Hadlum judged that 349 days was an outlier
- The question is what process gave rise to it:
  - An unusually long gestation?
  - Adultery (asserted by Mr. Hadlum)

Hadlum vs. Hadlum (1949)

- Average human gestation period is 280 days
- Mr. Hadlum judged that 349 days was an outlier
- The court ruled that the observation was **valid**, if extreme
- Mr Hadlum claimed that the observation was a **contaminant**
  - He did not want the observation to be **rejected**
  - He wanted it to be **identified**, with appropriate consequences

Introduction
Definition of an Outlier

> An outlying observation, or 'outlier', is one that appears to deviate markedly from other members of the sample in which it occurs.
>
> —F.E.Grubbs
> *Procedures for detecting outlying observations in samples* (1969)

> We shall define an outlier in a set of data to be an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data.
>
> —V.Barnett and T.Lewis
> *Outliers in Statistical Data* (1978)

Introduction
Anomalies as Items of Interest

- Anomalous credit card data $\rightarrow$ bank fraud/identity theft
- Anomalous network traffic $\rightarrow$ cyber-intrusion
- Anomalous MRI image $\rightarrow$ malignant tumour
- Anomalous readings from a spacecraft sensor $\rightarrow$ fault in component of spacecraft
- Anomalous astronomical data $\rightarrow$ a black hole
- Anomalous bump on LHC data $\rightarrow$ a new particle

Things to Consider

How to Find Anomalies

Naïve Approach

- Define a region representing normal behaviour. Declare anything outside that region as an anomaly.

This is difficult because:

- Defining the normal region is difficult
- Normal behaviour may not be static
- What is considered normal/anomalous may depend on the domain

Things to Consider

How to Find Anomalies

Naïve Approach

- Define a region representing normal behaviour. Declare anything outside that region as an anomaly.

This is difficult because:

- Defining the normal region is difficult
- Normal behaviour may not be static
- What is considered normal/anomalous may depend on the domain

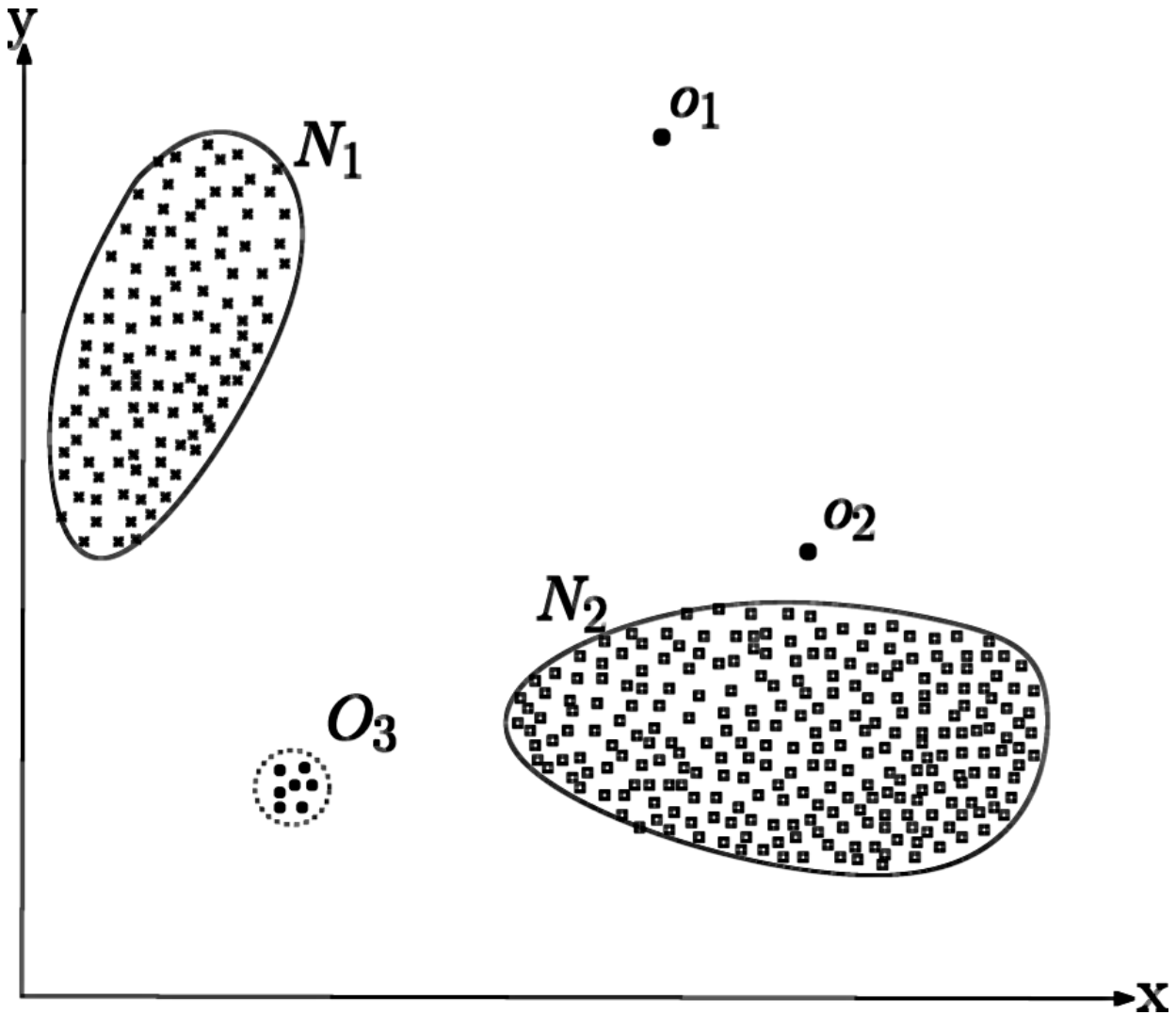
Things to Consider

Nature of the Input Data

- Data instance is usually a vector
- Data can have different characteristics:
  - Univariate
  - Multivariate
  - Attribute type can be binary/categorical/ordinal/ continuous
  - Multivariate data can have attributes of more than one type
  - Structured data: temporal, spatial, sequence, graph

Things to Consider

Type of Anomaly



Things to Consider

How to Handle Anomalies

- Inclusion
- Rejection
- Accommodation
- Identification

Things to Consider

Supervised/Semi-supervised/Unsupervised Scenarios

Supervised Approaches

- Requires training data for normal and anomaly classes
- Build a classifier (predictive model) for all classes
- Unseen data is compared to the model and classified as normal or anomalous
- Usually, anomalous instances are far fewer than normal instances
- Difficult to obtain accurate class labels for the anomaly instances

Things to Consider

Supervised/Semi-supervised/Unsupervised Scenarios

Semi-supervised Approaches

- Requires training data for the normal class only
- Build a one-class classifier
- Unseen data is compared to the model and classified as belonging to the normal class, or not
- Difficult to ensure that training data does not contain any anomalies

Things to Consider

Supervised/Semi-supervised/Unsupervised Scenarios

Unsupervised Approaches

Assumption

Normal instances are much more common than anomalous instances

- Does not require any training data
- Difficult to determine the threshold between normal data and outliers

Things to Consider

Labelling vs. Scoring

Labelling

- Binary output
- Data objects are labeled either as normal or outlier

Things to Consider

Labelling vs. Scoring

Scoring

- Continuous output, representing the degree of "outlierness"
- For each object an outlier score is computed
- Data objects can be sorted and ranked according to their scores
- Many scoring approaches focus on determining the top-n outliers
- Convert scores to labels by thresholding on the score

Things to Consider

Global vs. Local Anomalies

Global Outliers

Assumption

There is only one mechanism that generates the normal data instances

- Reference set is all other data objects
- Other outliers are also in the reference set and may distort the results

Things to Consider

Global vs. Local Anomalies

Local Outliers

- No assumption on the number of normal mechanisms
- Reference set is a small subset of data objects
- Resolution of the reference set can vary from a single object (local) to the entire database (global)
- Main problem: how to choose the optimal reference set

Overview

- Introduction
- Global Anomaly Detection
  - Statistical Approaches
  - Classification-based Approaches
  - Clustering-based Approaches
- Local Anomaly Detection
  - Distance-based Approaches
  - Density-based Approaches
- Anomaly Detection in High-dimensional Data
- Other Approaches

Global Anomaly Detection

- Statistical Approaches
  - Parametric
  - Nonparametric
- Classification-based Approaches
- Clustering-based Approaches

Global Anomaly Detection Approaches

Statistical Approaches

Assumption

Normal data instances occur in high-probability regions of a stochastic model; anomalies occur in the low-probability regions of the model.

- **Training:** Fit a statistical model to the training data
- **Testing:** Apply a statistical inference test to determine if an unseen instance belongs to the model

Statistical Approaches

Parametric Methods

- Normal observations $\mathbf{x}$ are generated by a process which is modelled as a parametric distribution
- Model is a Probability Density Function (PDF) $f(\mathbf{x}, \Theta)$
- Parameters Θ are estimated from training data
- **Labelling:** Statistical hypothesis test. Null hypothesis H_0 is that x was generated using the estimated distribution $f(\mathbf{x}, \Theta)$. If the statistical test rejects H_0, x is an anomaly.
- **Scoring:** Anomaly score for instance $x \in \mathbf{x}$ is the inverse of the PDF $f(x, \Theta)$

Statistical Approaches

Parametric Methods

- Normal observations $\mathbf{x}$ are generated by a process which is modelled as a parametric distribution
- Model is a Probability Density Function (PDF) $f(\mathbf{x}, \Theta)$
- Parameters Θ are estimated from training data
- **Labelling:** Statistical hypothesis test. Null hypothesis H_0 is that x was generated using the estimated distribution $f(\mathbf{x}, \Theta)$. If the statistical test rejects H_0, x is an anomaly.
- **Scoring:** Anomaly score for instance $x \in \mathbf{x}$ is the inverse of the PDF $f(x, \Theta)$

Univariate Models

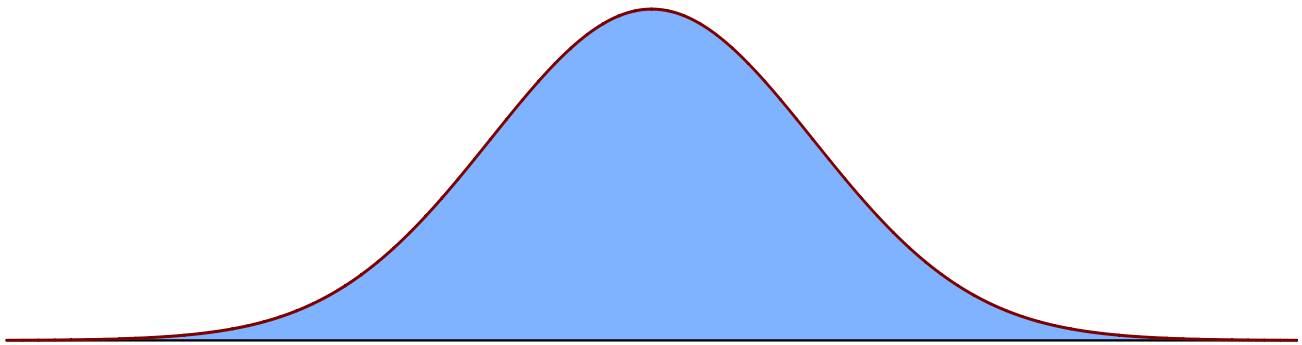
Warning: Gaussian Distributions Ahead

Parametric Methods

Gaussian Model

Assumption

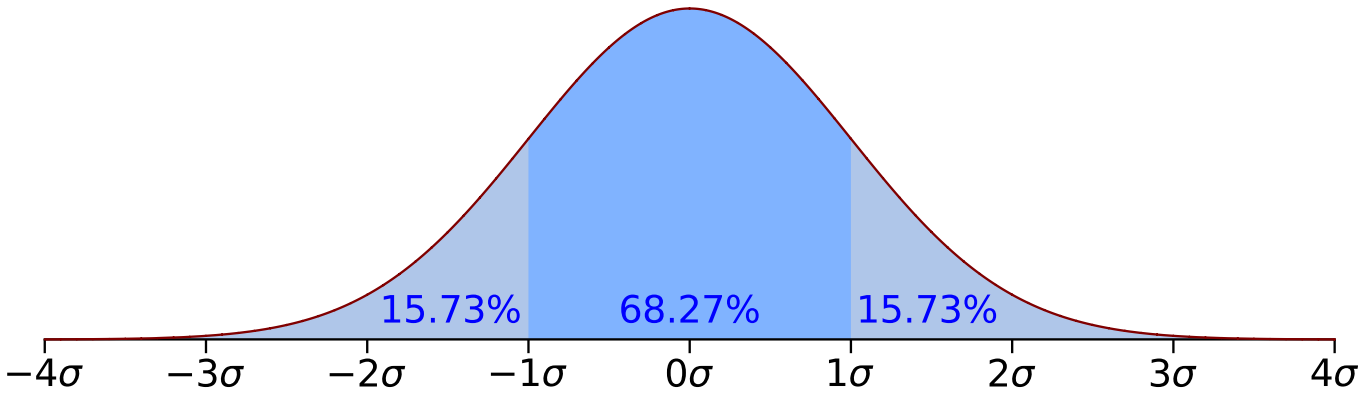
Data is generated by a Gaussian process. Therefore it is normally distributed about the mean.



Univariate Models

Gaussian Model

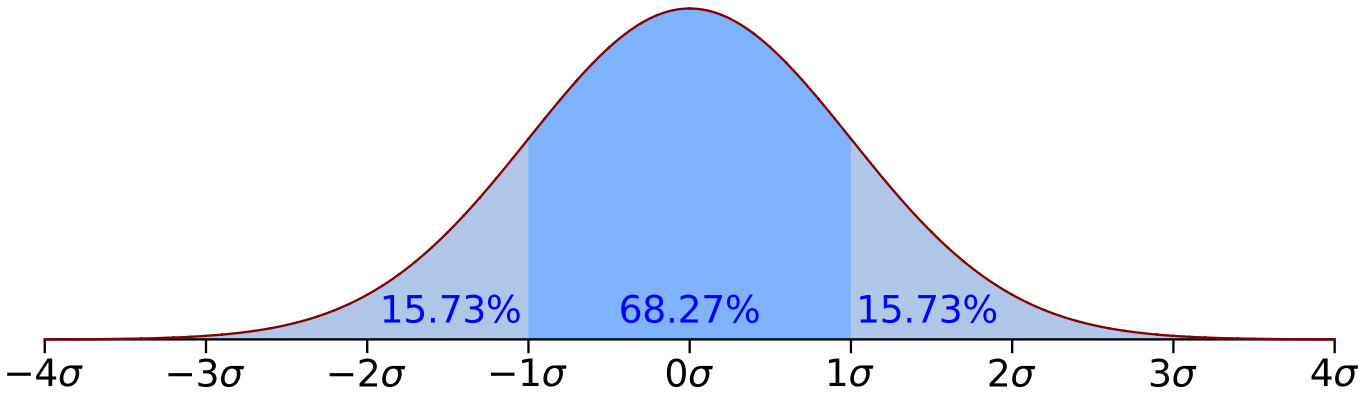
- Assume a normal distribution $\eta(\mu, \sigma^2)$ with mean μ and variance σ^2
- Estimate parameters $\Theta = \{\mu, \sigma\}$
- **Labelling:** Choose a threshold, *e.g.* 3σ. Everything outside the normal range ($\mu \pm 3\sigma$) is an outlier.
- **Scoring:** Distance of data to the mean μ



Univariate Models

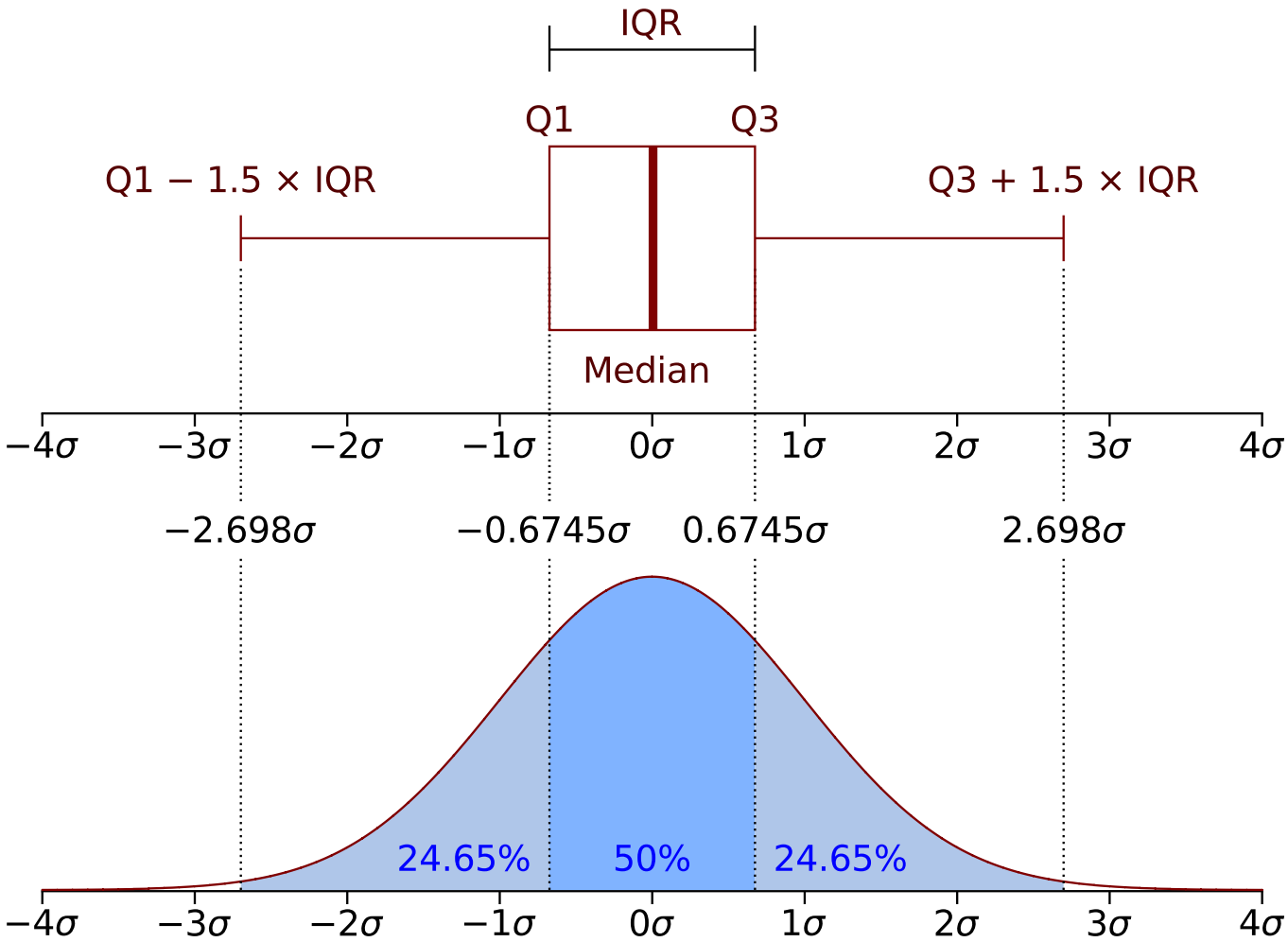
Gaussian Model

- Assume a normal distribution $\eta(\mu, \sigma^2)$ with mean μ and variance σ^2
- Estimate parameters $\Theta = \{\mu, \sigma\}$
- **Labelling:** Choose a threshold, *e.g.* 3σ. Everything outside the normal range ($\mu \pm 3\sigma$) is an outlier.
- **Scoring:** Distance of data to the mean μ



Univariate Models

Box Plot Rule

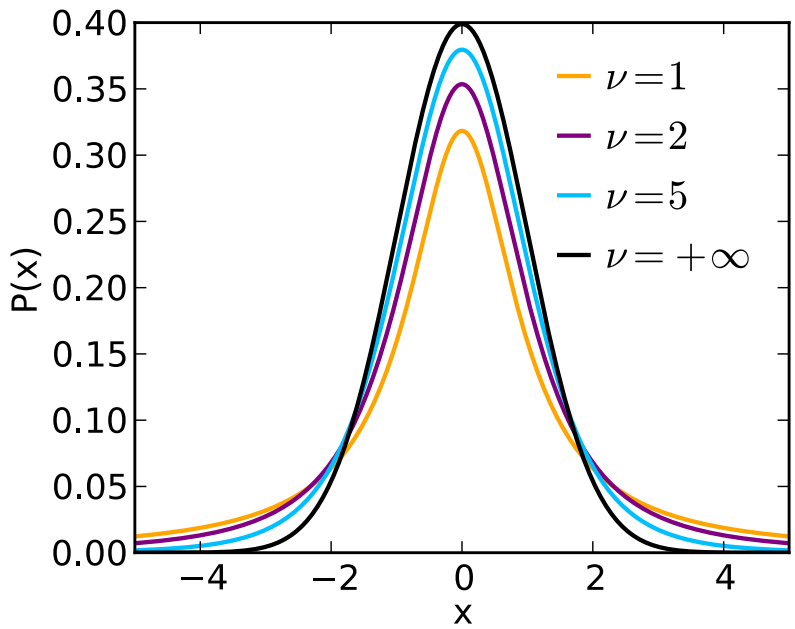


Univariate Models

Student's *t*-test

Univariate Models

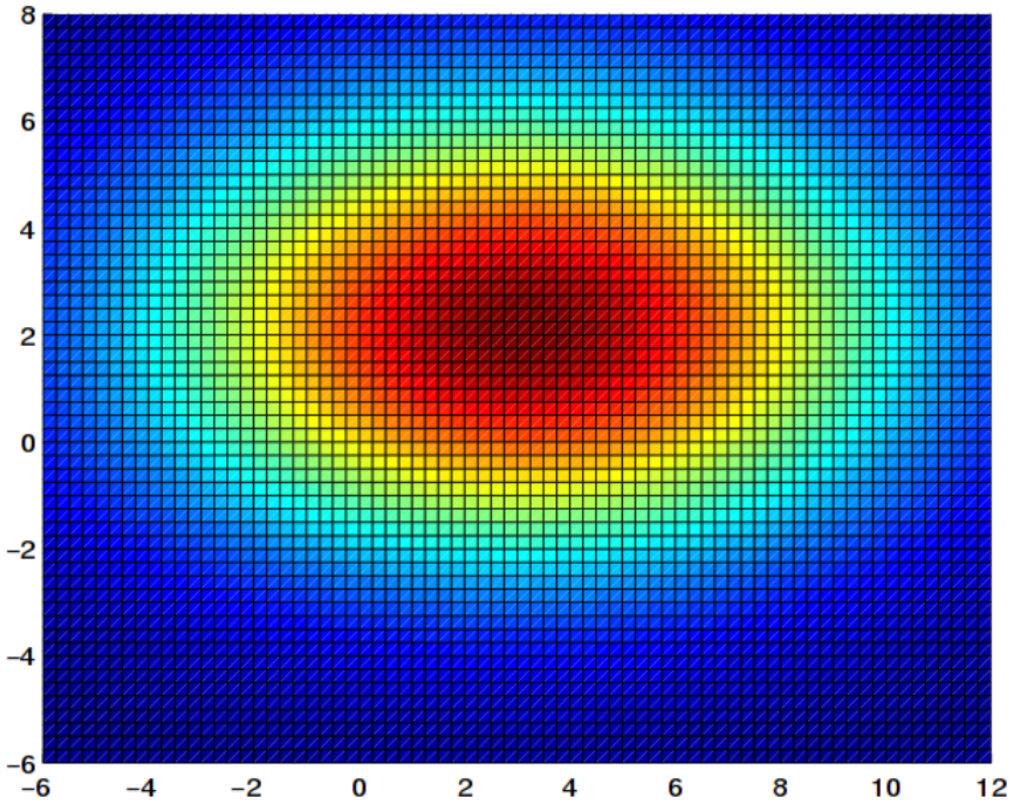
Student's *t*-test



- ***t*-distribution:** distribution of the location of the true mean relative to the sample mean
- ***t*-test:** statistical hypothesis test; estimates confidence that test sample comes from the same distribution as training samples
- If H_0 is rejected, test sample is an anomaly
- Suitable for small sample sizes where normal behaviour can be easily quantified

Multivariate Models

Multivariate Gaussian Distribution



- Probability Density Function (PDF)

$$P = \eta(\mu, \Sigma)$$

- Mean

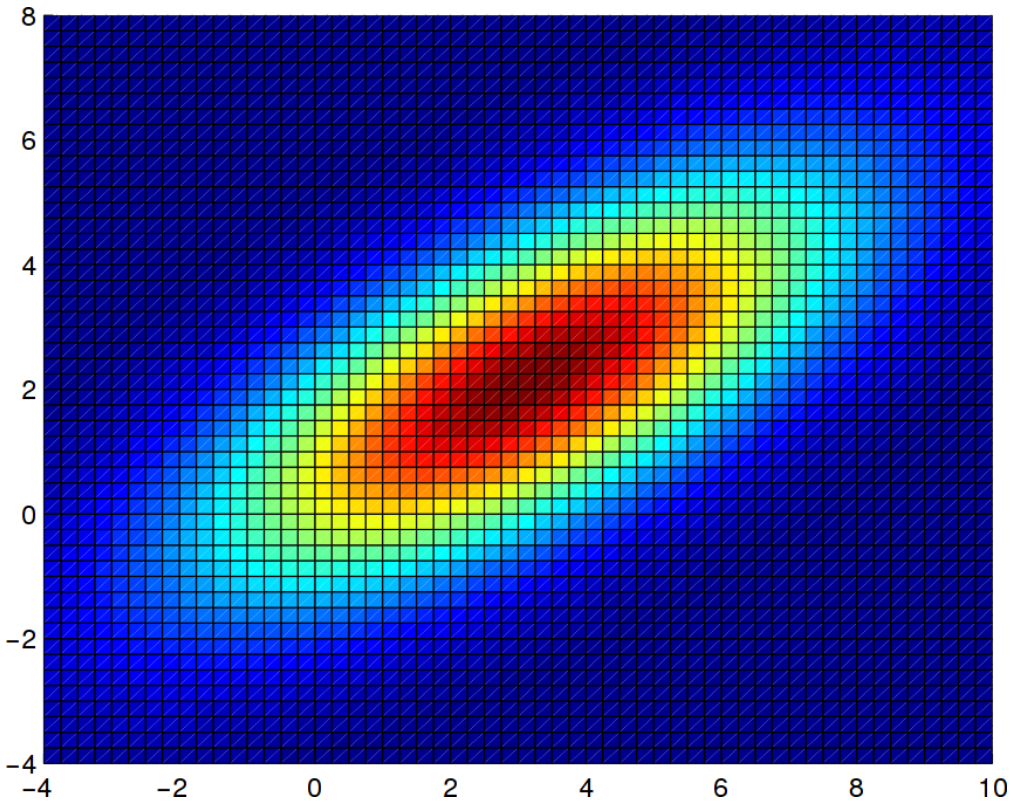
$$\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- Covariance

$$\Sigma = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}$$

Multivariate Models

Multivariate Gaussian Distribution



- Probability Density Function (PDF)

$$P = \eta(\mu, \Sigma)$$

- Mean

$$\mu = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

- Covariance

$$\Sigma = \begin{bmatrix} 10 & 5 \\ 5 & 5 \end{bmatrix}$$

Multivariate Models

Mahalanobis Distance

- **Mahalanobis Distance:** measures the distance between a point $\mathbf{p}$ and a probability distribution $P = \eta(\mu, \Sigma)$:

$$d(\mathbf{p}, P) = \sqrt{(\mathbf{p} - \mu)^T \cdot \Sigma^{-1} \cdot (\mathbf{p} - \mu)}$$

- Distance is zero if $\mathbf{p}$ is at the mean of P
- Distance grows as $\mathbf{p}$ moves away from the mean along each principal component axis
- Multi-dimensional generalisation of measuring how many standard deviations $\mathbf{p}$ is from the mean of P.

Multivariate Models

Mahalanobis Distance

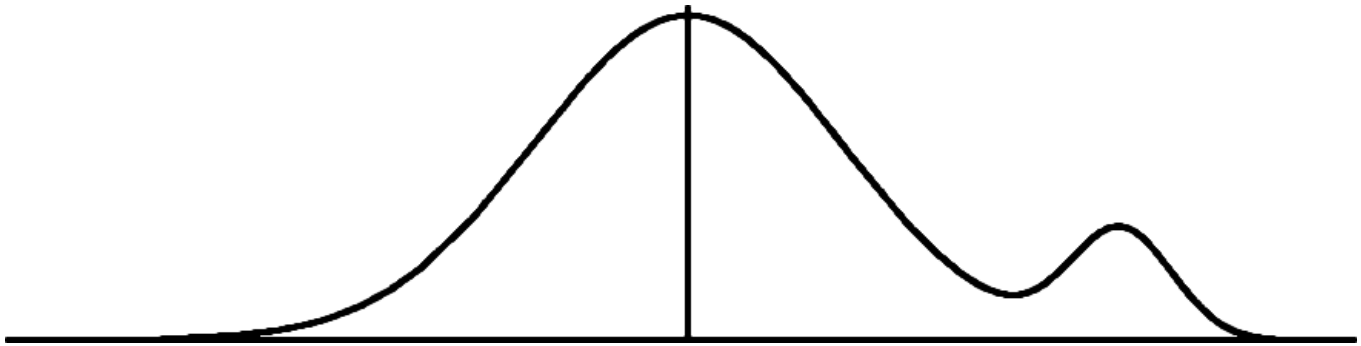
- **Mahalanobis Distance:** measures the distance between a point $\mathbf{p}$ and a probability distribution $P = \eta(\mu, \Sigma)$:

$$d(\mathbf{p}, P) = \sqrt{(\mathbf{p} - \mu)^T \cdot \Sigma^{-1} \cdot (\mathbf{p} - \mu)}$$

- If axes are rescaled to unit variance, Mahalanobis distance corresponds to standard Euclidean distance
- Mahalanobis distance is thus unitless and scale-invariant, and takes into account the correlations of the data set

Mixture Models

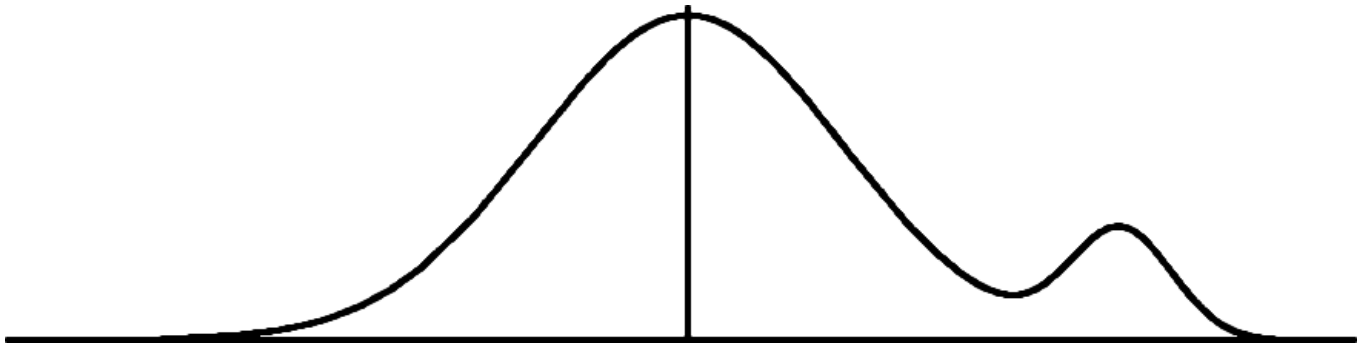
The Bombes at Bletchley Park



"You'll notice that the girls who are assigned to that particular duty are unusually tall. If the Germans were to somehow get their hands on the personnel records for all of the people who work at Bletchley Park, and graph their heights on a histogram, they would see a normal bell-shaped curve, representing most of the workers, with an abnormal bump on it—representing the unusual population of tall girls who we have brought in to work the plug boards."

Mixture Models

The Bombes at Bletchley Park



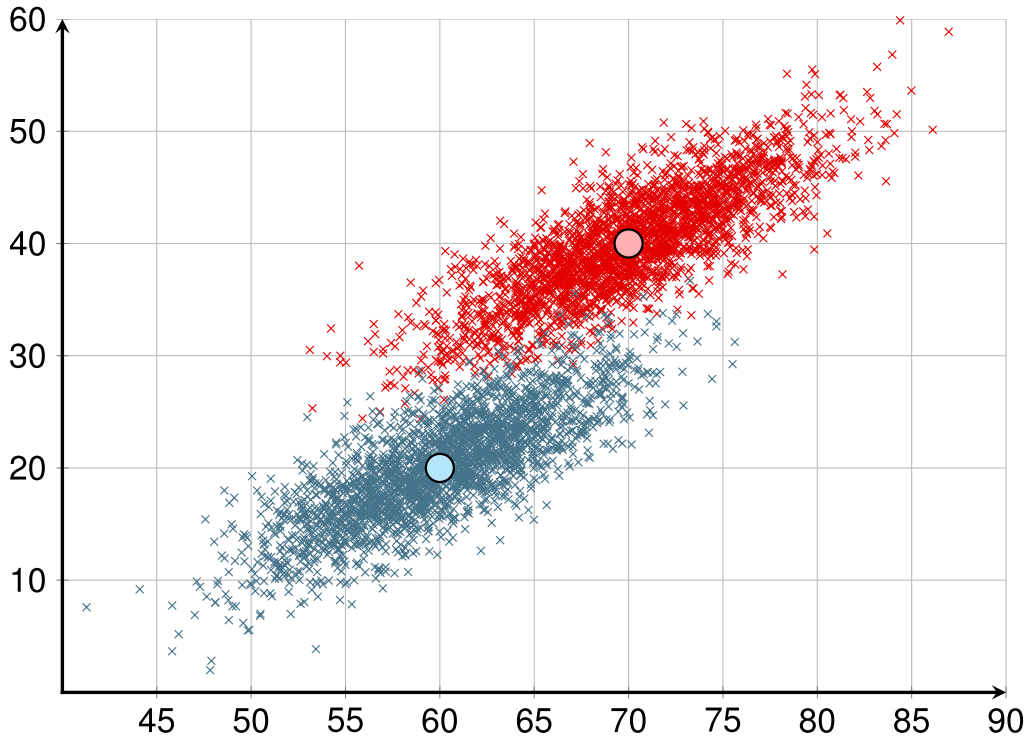
"Yes, I see." Waterhouse says, "and someone like Rudy—Dr. von Hacklheber—would notice the anomaly, and wonder about it."

— Neal Stephenson,
The Cryptonomicon

Mixture Models

Gaussian Mixture Model

- A **Mixture Model** is a mixture of parametric statistical distributions
- Most typically, the Gaussian Mixture Model (GMM)



Mixture Models

Gaussian Mixture Model

- GMM is a mixture of N independent Gaussian distributions
- The contribution of each Gaussian to the mixture is determined by a weight vector ϕ

$$f(\mathbf{x}, \Theta) = \sum_{i=1}^{N} \phi_i \cdot \eta(\mu_i, \Sigma_i)$$

Mixture Models

Anomaly Detection by Expectation Maximisation

$$\mathbf{D} = \lambda \mathbf{A} + (1 - \lambda)\mathbf{M}$$

- Assume a GMM with two Gaussian mixtures:
  - $\mathbf{M}(\mu_1, \Sigma_1)$ for normal data
  - $\mathbf{A}(\mu_2, \Sigma_2)$ for anomalies
- $\mathbf{D}$ is the actual probability distribution of the entire data
- λ is the prior probability that a data point is an anomaly
- Initially all points are in $\mathbf{M}$

Mixture Models

Anomaly Detection by Expectation Maximisation

$$\mathbf{D} = \lambda\mathbf{A} + (1 - \lambda)\mathbf{M}$$

- **Expectation (E) Step:** Each point is assigned a probability of being in $\mathbf{A}$ based on how much the distributions change if the point is removed from $\mathbf{M}$ and added to $\mathbf{A}$
- **Maximisation (M) Step:** Points are assigned to $\mathbf{A}$ or $\mathbf{M}$ based on the (log-)likelihood function calculated in the E-step
- Iterate until we converge on the **Maximum Likelihood Estimate (MLE)**, the parameters which provide the highest-probability explanation of the data

Mixture Models

Anomaly Detection by Expectation Maximisation

- **Supervised Mode:** model normal and anomalous instances as separate mixtures
- **Semi-supervised Mode:** model only normal instances ("background") as mixtures
- Can be generalised to an arbitrary number of Gaussian mixtures
- Variational Bayesian algorithms can produce better results, *e.g.* Dirichlet Process Gaussian Mixture Model (DPGMM)

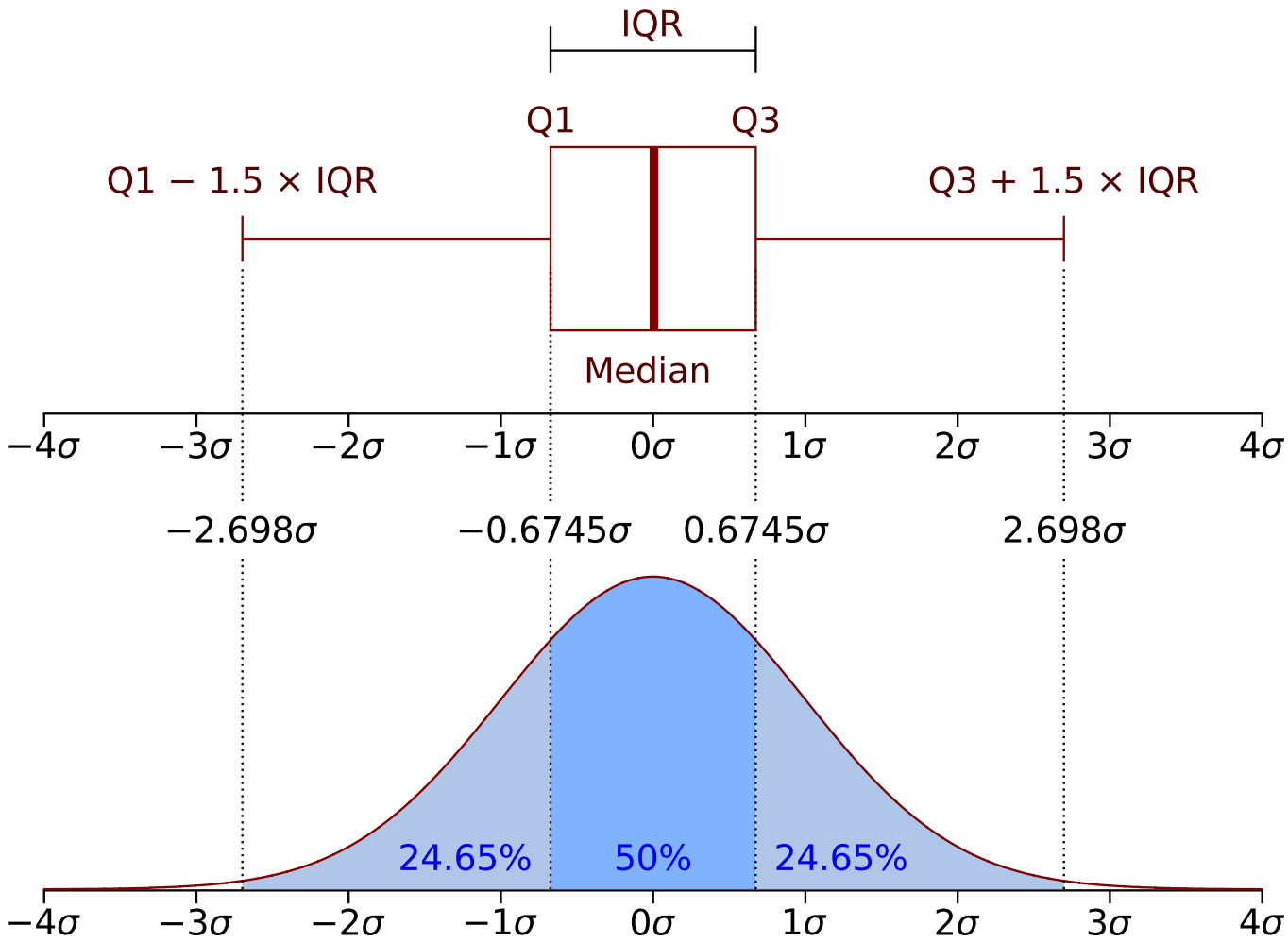
Statistical Approaches

Non-parametric Methods

- Model structure is not defined *a priori*, it is determined from given data.
- More suitable where the underlying distribution of data is unknown, as it makes make fewer assumptions.

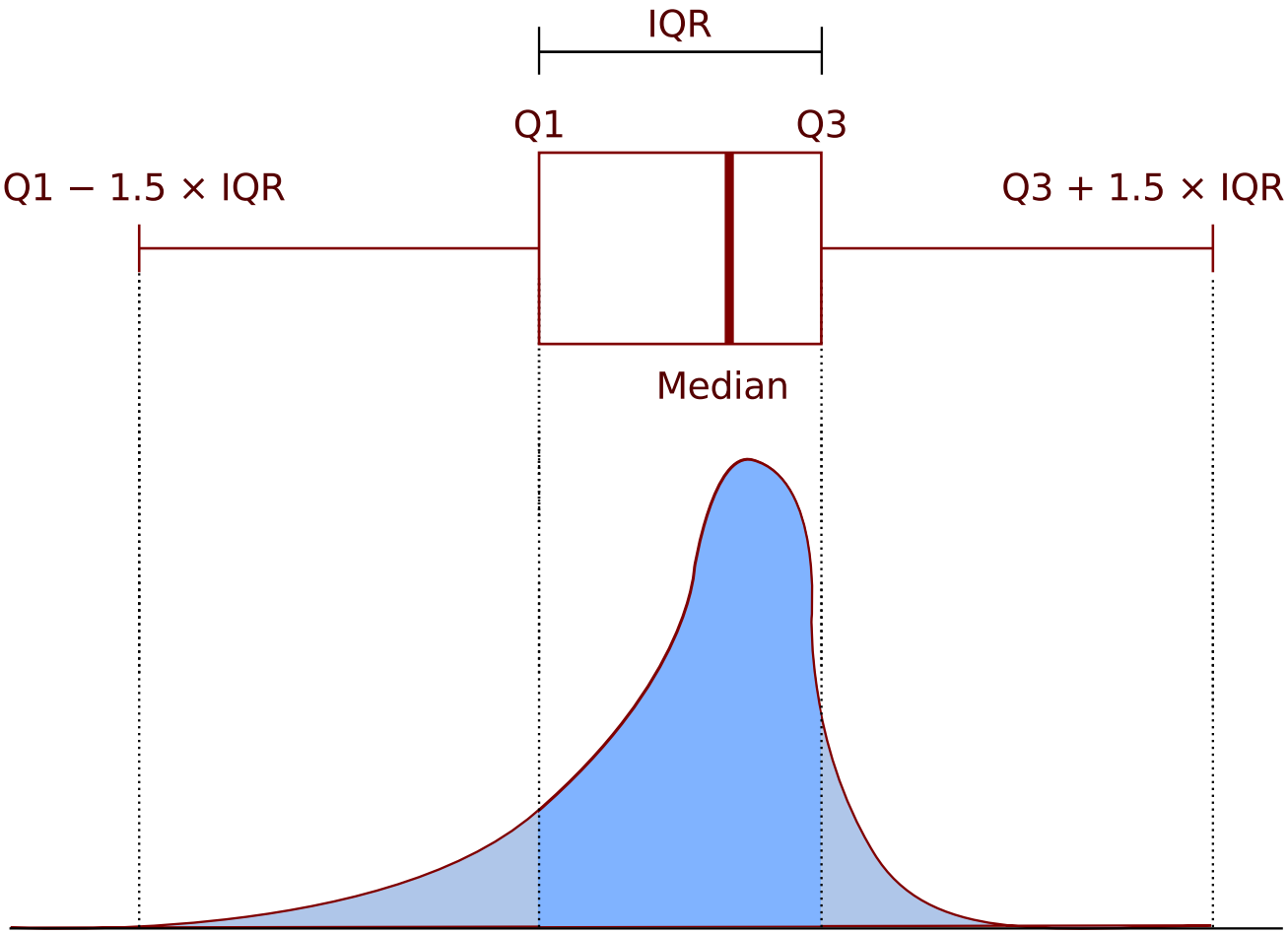
Nonparametric Methods

Box Plot Rule



Nonparametric Methods

Box Plot Rule



Nonparametric Methods

Histogram-based Approach

Anomaly detection at CERN experiments

Yesterday I got a brilliant idea on how to implement automatic anomaly detection at CERN experiments. Today this work is done manually—many students/PhD students are looking at different distributions online. This is quite unreliable, and it's quite expensive, since you need many people to work all the time (nobody is paid for this—but you anyway spend money on travels). So, the basic idea is quite simple: one can bin each variable and look at distributions within each of bins. Knowing, that number of events observed inside each bin is Poisson-distributed, one can detect anomalies.

Nonparametric Methods

Histogram-based Approach

- Model is based on counting frequency of data assigned to pre-defined bins
- **Semi-supervised:** create a model of normal data only
- **Labelling:** If test data falls into a bin defined during training, it is normal, otherwise anomalous
- **Scoring:** Assign an anomaly score based on the height of the bin that test data is assigned to

Statistical Approaches

Computational Complexity

- Strongly depends on the statistical model chosen
- Fitting single parametric distributions is typically linear in data size and number of attributes
- Fitting complex models using iterative techniques like EM are typically linear for each iteration but may be slow to converge

Statistical Approaches

Advantages

- If the assumptions about the underlying model hold true, gives a statistically justifiable model for anomaly detection
- Anomaly score is associated with a confidence interval, which allows scoring of anomalies
- If the distribution estimation is robust to anomalies, statistical methods can be used in an unsupervised mode

Statistical Approaches

Disadvantages

- Assumptions may not hold true
- Not always obvious which test statistic is best
- Difficult to construct a hypothesis test for complex high-dimensional data
- Histogram-based techniques are easy to implement but cannot capture interactions between different attributes

Classification-based Approaches

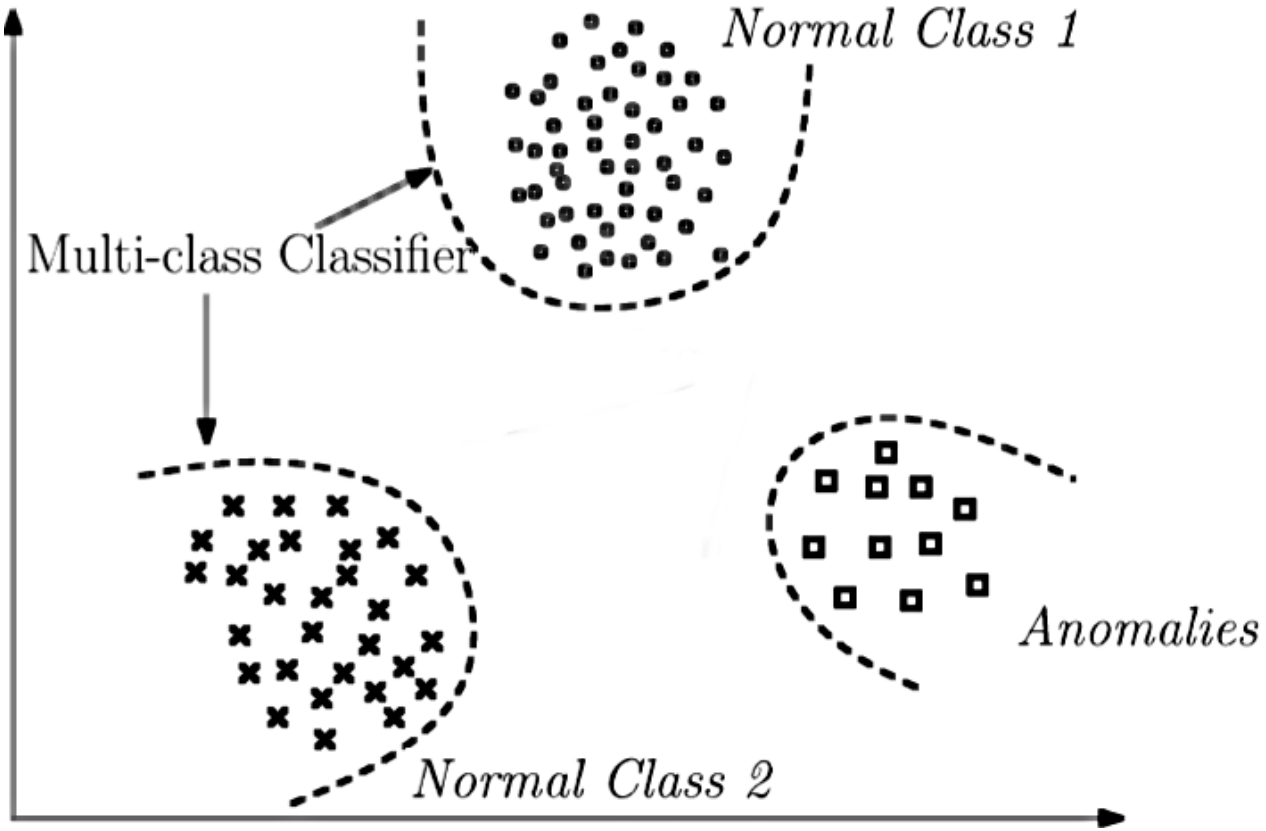
Assumption

A classifier that can distinguish between normal and anomalous classes can be learned in the given feature space

- **Training:** Learn a model (classifier) from a set of labelled data instances
- **Testing:** Classify a test instance into one of the classes using the learned model

Classification-based Approaches

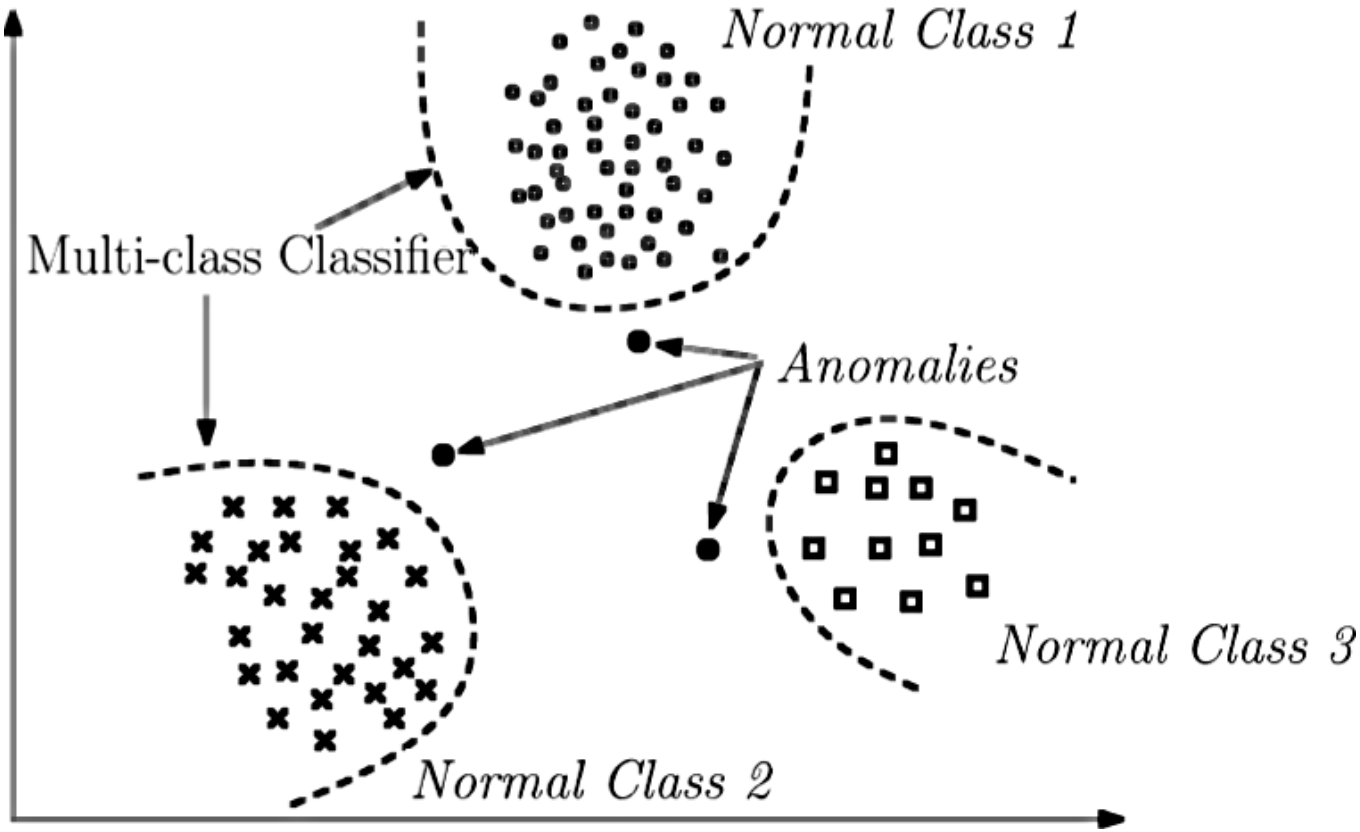
Supervised Multi-class Classification



- Training data contains instances belonging to one or more normal classes and the anomalous class

Classification-based Approaches

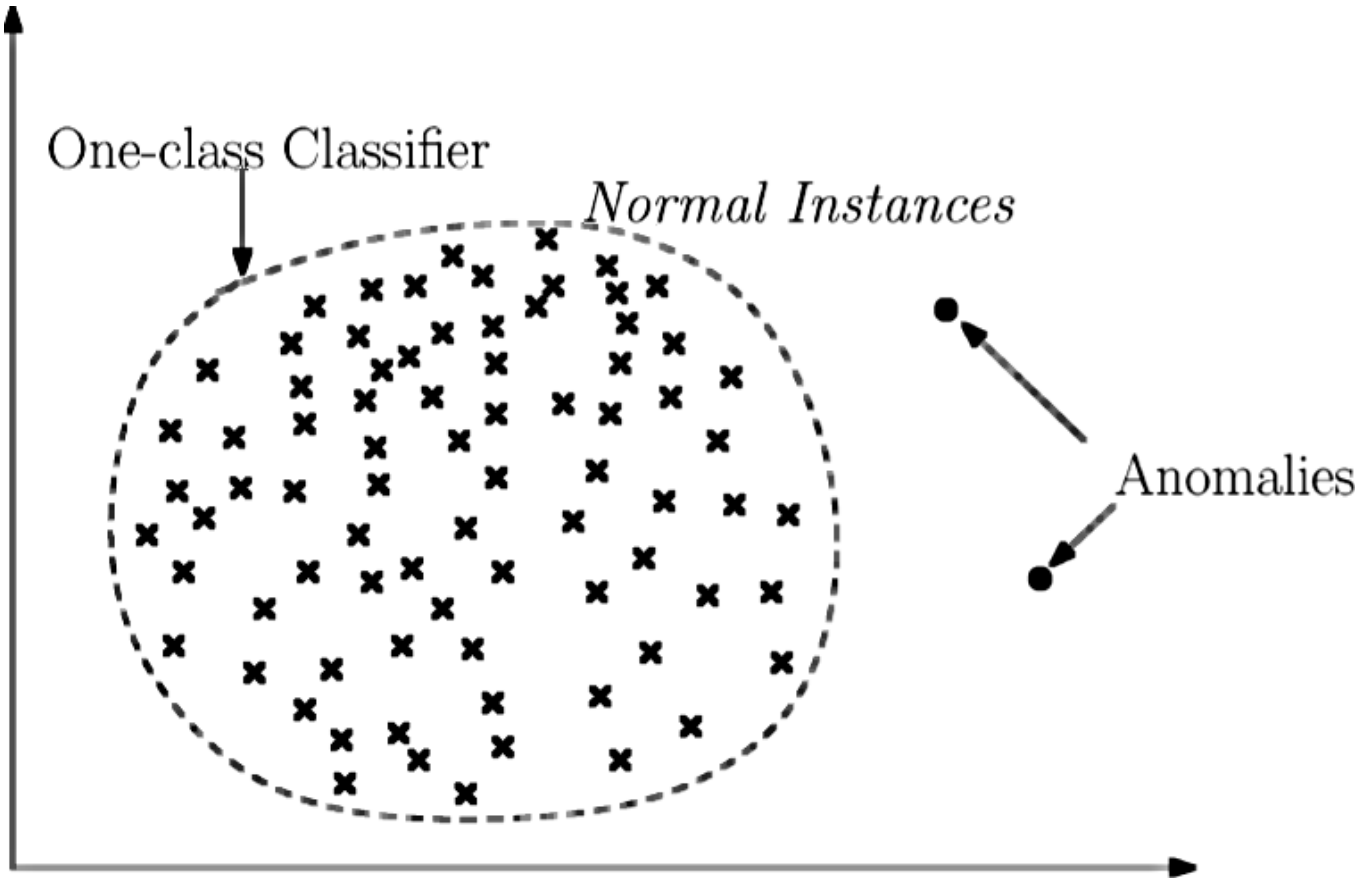
Semi-supervised Multi-class Classification



- Training data contains instances belonging to normal classes only

Classification-based Approaches

One-class Classification



- A single boundary is learned for all normal data

Classification-based Approaches

Rule-based Classifiers

- Learn rules from training data:

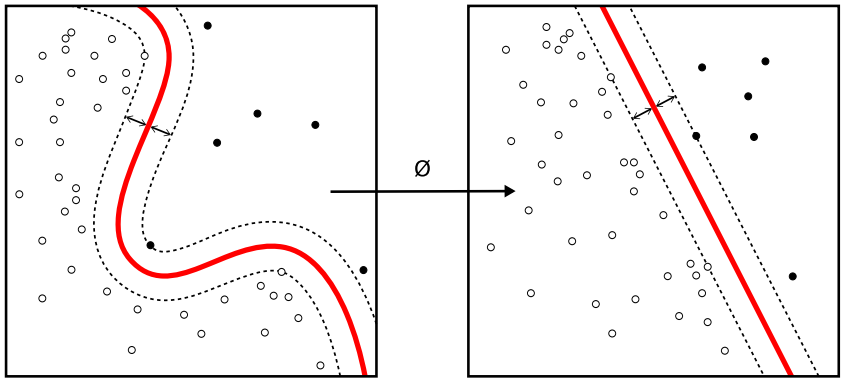
$$(A_1 = v_1 \wedge A_2 = v_2 \ldots) \implies C_1$$

- Rule quality is evaluated by:
  - Support/Coverage
  - Confidence/Accuracy

- **Labelling:** test instances are matched to the best rule
- **Scoring:** take the inverse of the confidence score

Classification-based Approaches

Kernel Function-based Classifiers



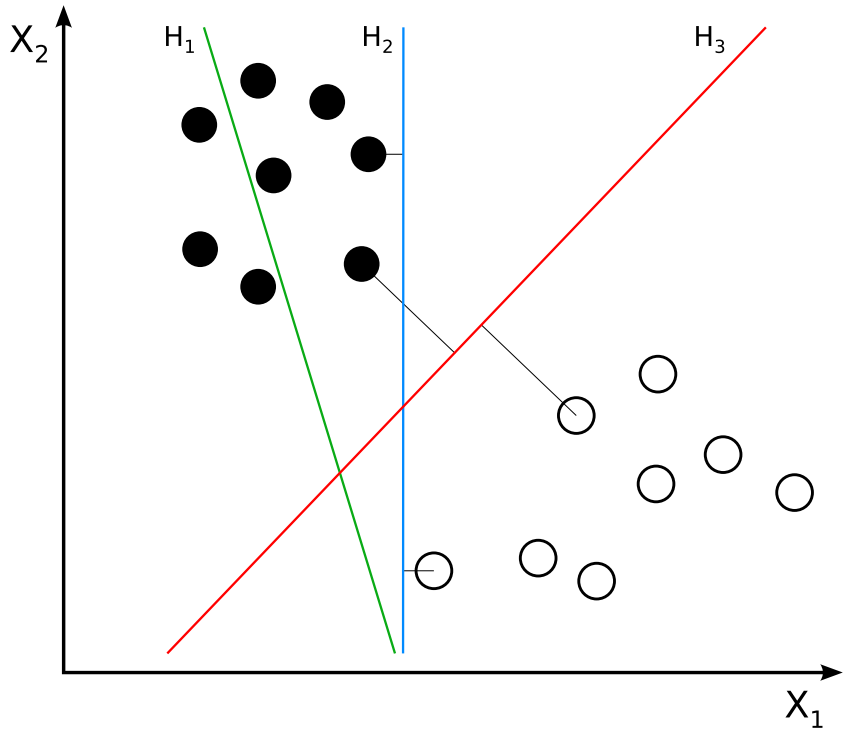
- Instance-based Learning
- **Kernel function:** a similarity function over pairs of data points

$$k : \chi \times \chi \implies \mathcal{R}$$

- Typically $\mathcal{O}(N^2)$ complexity

Classification-based Approaches

Support Vector Machine (SVM)



- Kernel is a Radial Basis Function (RBF)

$$\phi(\mathbf{x}, \mathbf{c}) = \phi(||\mathbf{x} - \mathbf{c}||)$$

- Points outside the learned boundary are anomalies

Classification-based Approaches

Bayesian Approaches

- Suitable for Categorical Data
- **One-class:** Naïve Bayes

$$Posterior = \frac{Likelihood \times Prior}{Evidence} \qquad P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- **Multi-class:** Bayesian Networks

Classification-based Approaches

Neural Networks

- **Multi-class:** Train network to accept/reject normal instances
- **One-class:** Replicator Neural Network
  - Create a multi-layer feed-forward neural network with same number of input/output neurons
  - Network reconstructs each data instance from the input
  - Reconstruction error is used as the anomaly score

Classification-based Approaches

Advantages

- Testing phase is fast, as items are compared against a precomputed model

Classification-based Approaches

Disadvantages

- Model accuracy depends on the availability of accurately-labelled training data
  - Training data usually will not cover all possible types of anomaly
  - Training data often has to be labelled manually
- Classifiers usually only offer labelling, not scoring

Clustering-based Approaches

K-Means Clustering

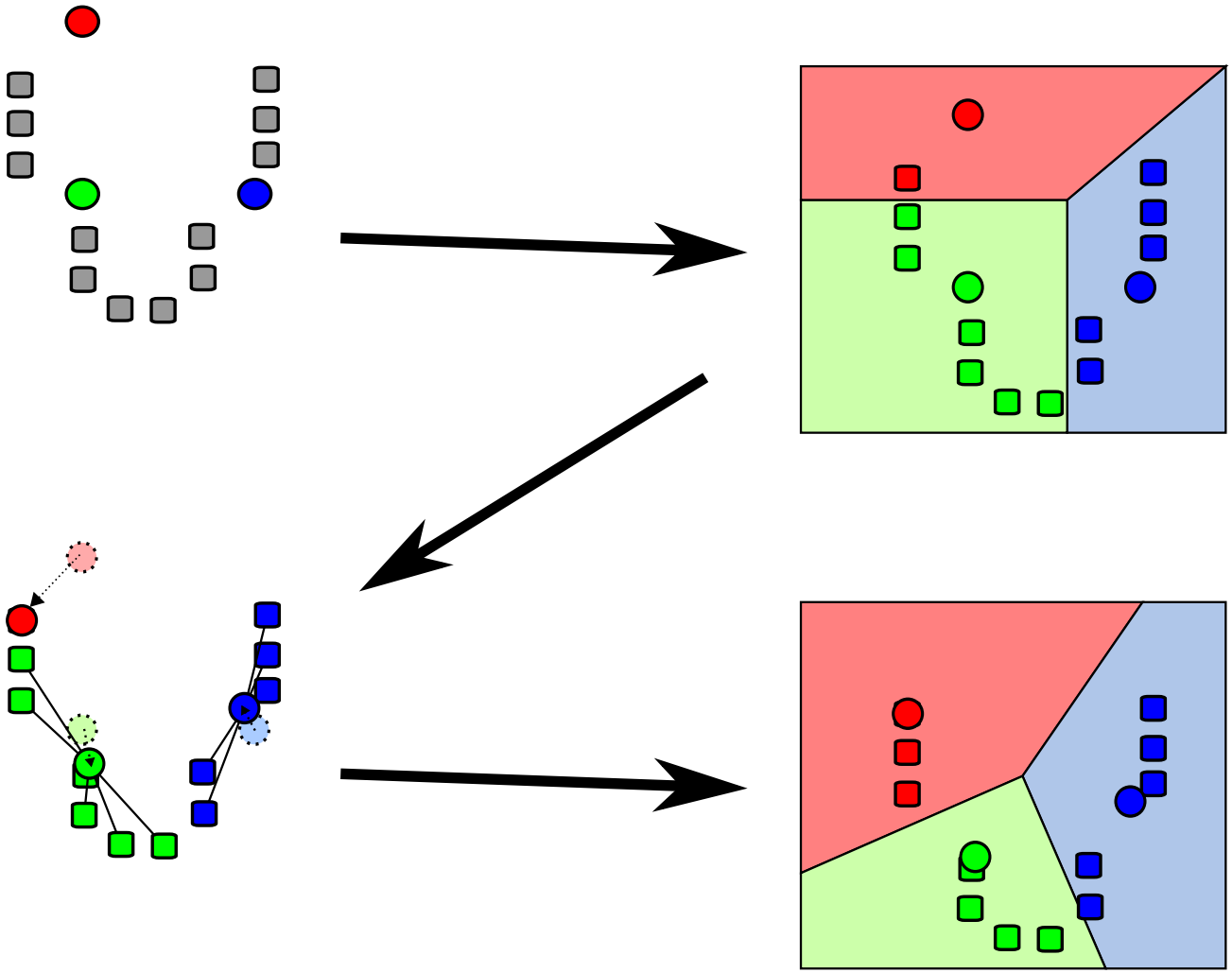
- Partition data into k clusters
- Each point belongs to the cluster with the nearest mean

$$\arg\min_{s} \sum_{i=1}^{k} \sum_{x \in S_i} ||\mathbf{x} - \mu_i||^2$$

- Parameter k is selected by the user
- Uses an iterative approach to converge to a local optimum

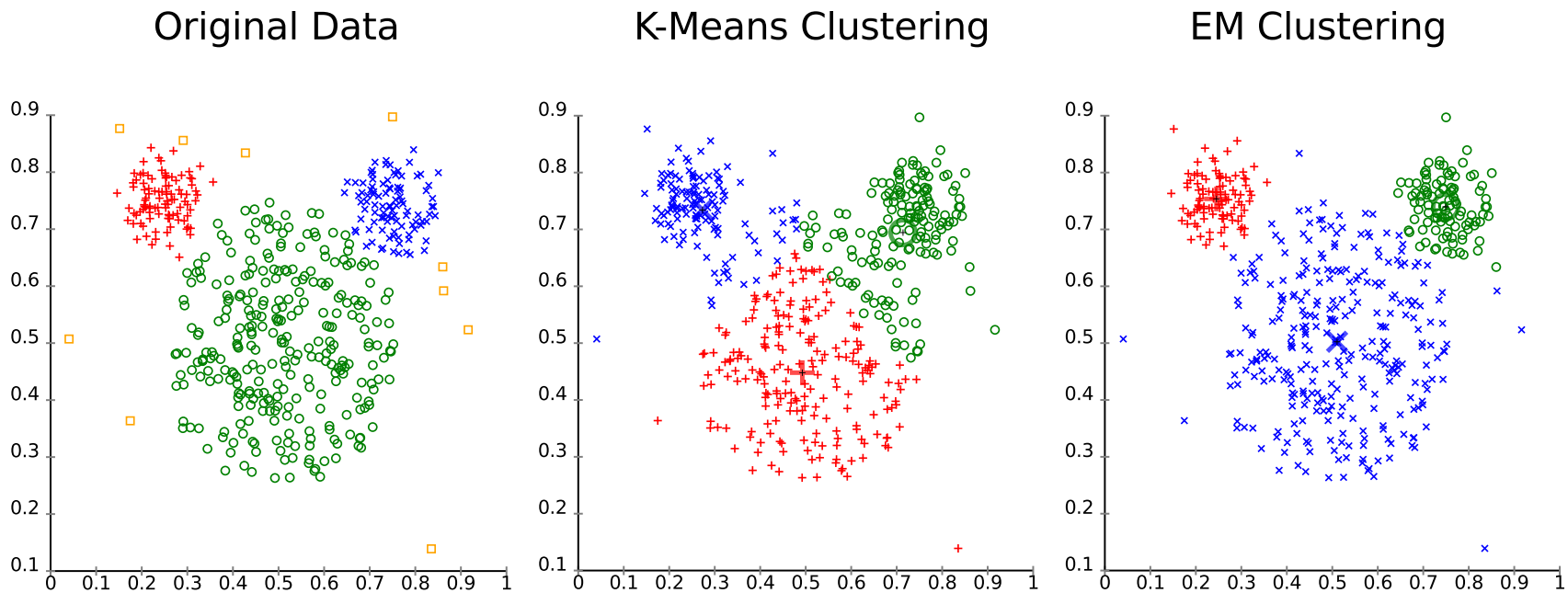
Clustering-based Approaches

K-Means Clustering



Clustering-based Approaches

EM Clustering



Clustering-based Approaches

Unsupervised Cluster-based Anomaly Detection

Assumption

Normal data instances belong to a cluster in the data. Anomalies do not belong to any cluster.

- Apply a known clustering algorithm to the dataset
- **Labelling:** Instances which do not belong to any cluster are declared as anomalies

Clustering-based Approaches

Semi-supervised Cluster-based Anomaly Detection

Assumption

Normal data instances lie close to their cluster centroid. Anomalies are far away from their cluster centroid.

- Apply a known clustering algorithm to the training data
- Compare the test data to the model of cluster centroids
- **Scoring:** the anomaly score for each data instance is the distance to its closest cluster centroid

Clustering-based Approaches

Local Anomaly Detection

Assumption

Normal data instances belong to large and dense clusters. Anomalies belong to small or sparse clusters.

- Useful to detect anomalies which form into clusters
- Cluster-based Local Outlier Factor (CBLOF): evaluates the size of the cluster and the distance to cluster centroid

Clustering-based Approaches

Advantages

- Can be applied to any data type for which a clustering algorithm exists
- In supervised mode, test phase is very fast (number of clusters is small compared to number of data points)
- Can operate in unsupervised mode

Clustering-based Approaches

Disadvantages

- Performance is highly dependent on how well the clustering algorithm captures the cluster structure of normal instances
- Techniques which detect anomalies as a byproduct of clustering are not optimised for anomaly detection
  - Some clustering algorithms force all points to be added to a cluster
  - Some techniques don't work if anomalies are clustered
- Some clustering algorithms have high $\mathcal{O}(N^2 d)$ computational complexity

Summary

- Introduction
- Global Anomaly Detection
  - Statistical Approaches
  - Classification-based Approaches
  - Clustering-based Approaches
- Local Anomaly Detection
  - Distance-based Approaches
  - Density-based Approaches
- Anomaly Detection in High-dimensional Data
- Other Approaches

Summary

- Introduction
- Global Anomaly Detection
  - Statistical Approaches
  - Classification-based Approaches
  - Clustering-based Approaches
- Local Anomaly Detection
  - Distance-based Approaches
  - Density-based Approaches
- Anomaly Detection in High-dimensional Data
- Other Approaches