Outliers detection

Machine Learning

25.05.2018

Lecture plan

- Introduction into outliers
- Outliers types
- Outliers detection
- Anomaly detection
- Working with outliers

Lecture plan

- Introduction into outliers
- Outliers types
- Outliers detection
- Anomaly detection
- Working with outliers

Introduction

Outlier = distant point

Why it could happen?

- Measurement error
- Heavy-tail distribution
- Mixture of two distributions
- Systematic errors

Introduction

Outliers

- May include max or min or both or none
- Mean not always good for detection

Lecture plan

- Introduction into outliers
- Outliers types
- Outliers detection
- Anomaly detection
- Working with outliers

Outliers types

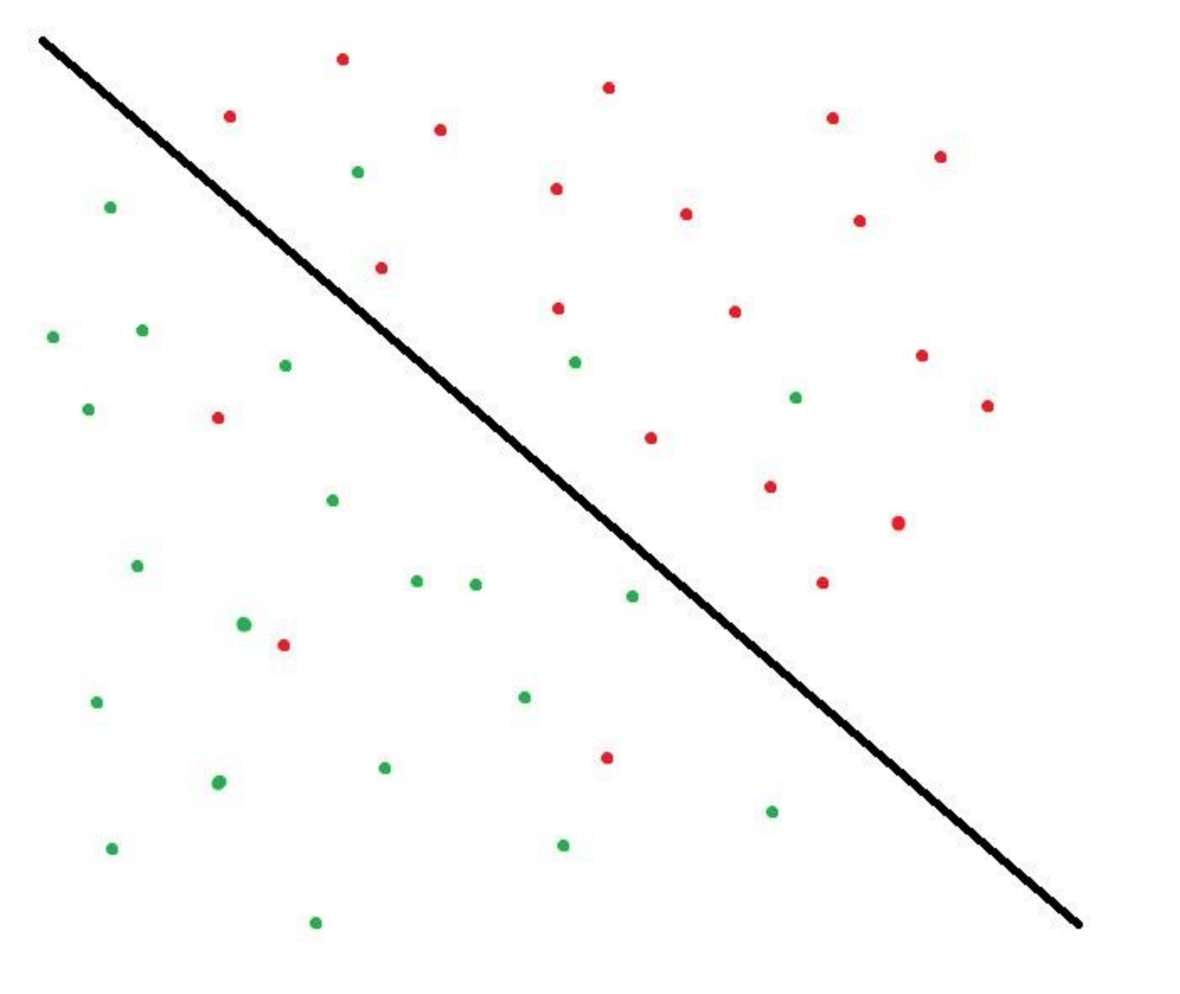
- Point
- Contextual
- Collective

Point outliers

Point outliers are the simplest ones:

Individual points that could be considered anomalous with respect to the rest of the data

Point outliers

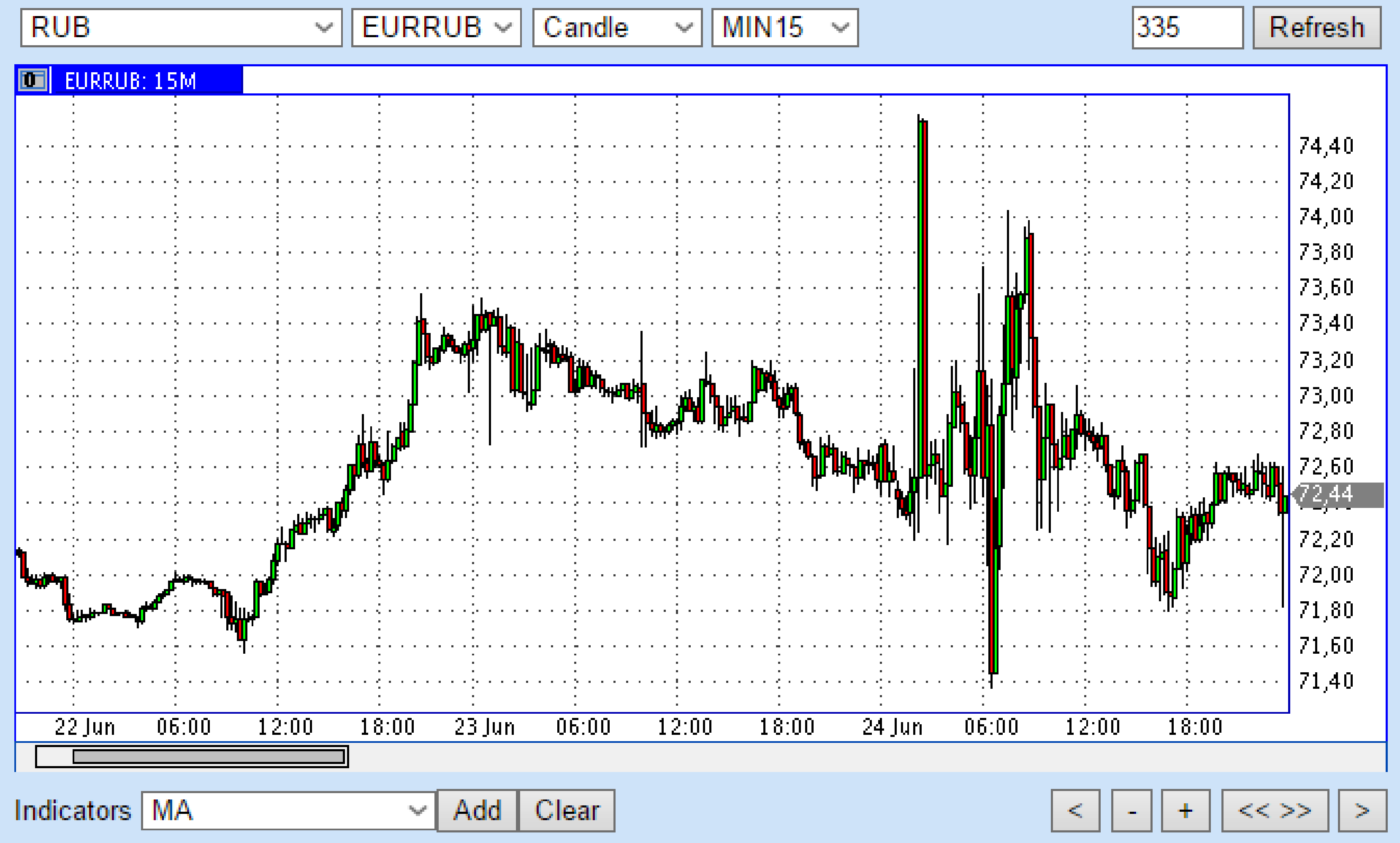


Contextual outliers

Contextual outliers have some specific context. Basically, attributes are divided into two parts:

- Contextual attributes – to determine context, for example time in time-series
- Behavioral attributes – to determine other specific parameters

Contextual outliers

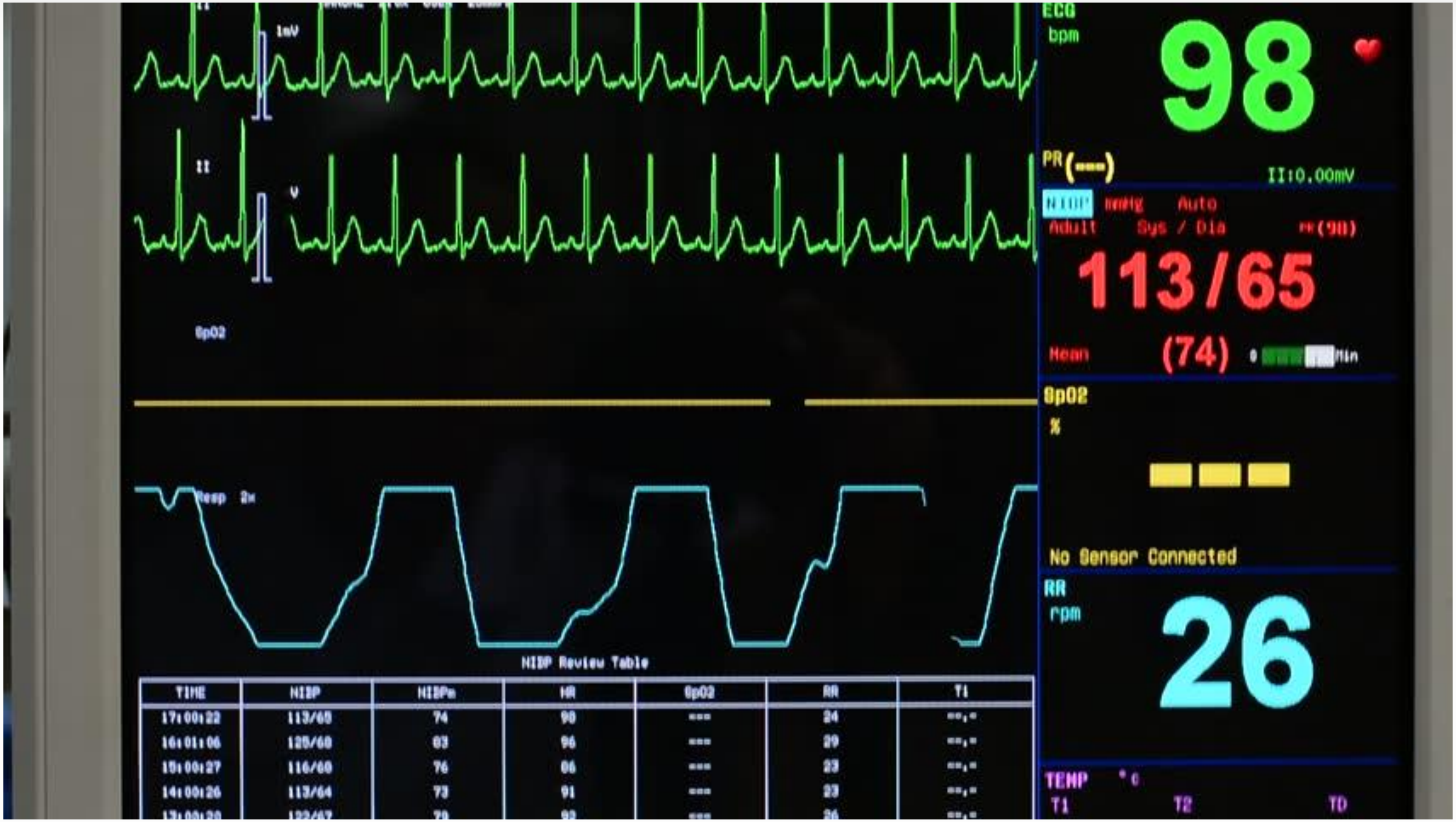


Collective outliers

Collective outliers – collection of anomalous data points with respect to entire dataset.

- May consist of several collective outliers
- Individual points may not be an outliers inside collective outliers

Collective outliers



NIBP Review Table

TIME	NIBP	NIBPm	HR	SpO2	RR	T1
17:00:22	113/65	74	98	---	24	--.-
16:01:06	125/60	83	96	---	29	--.-
15:00:27	116/60	76	86	---	23	--.-
14:00:26	113/64	73	91	---	23	--.-
13:00:20	122/67	79	92	---	26	--.-

Lecture plan

- Introduction into outliers
- Outliers types
- Outliers detection
- Anomaly detection
- Working with outliers

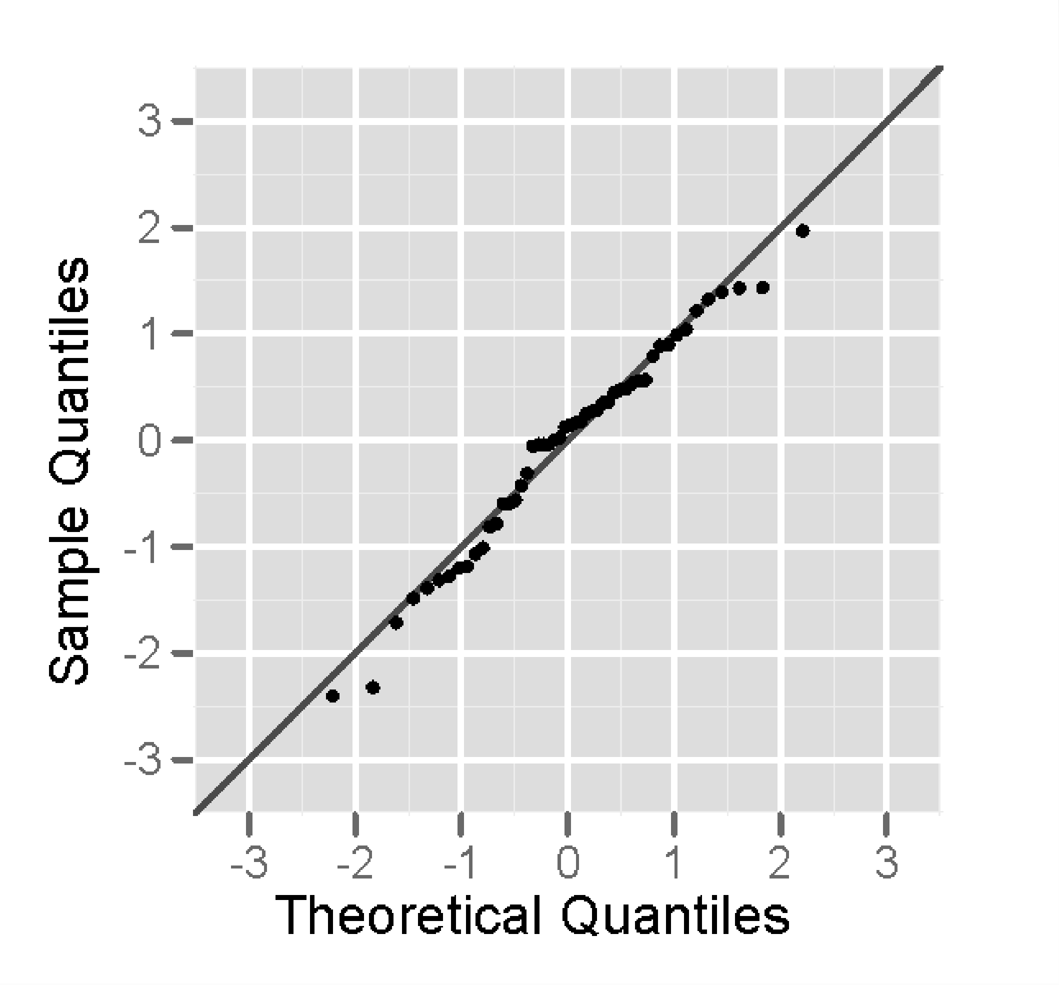
Outliers detection

Main approaches to outliers detection:

- Normal probability plots
- Model-based
- Box plots

Outliers detection

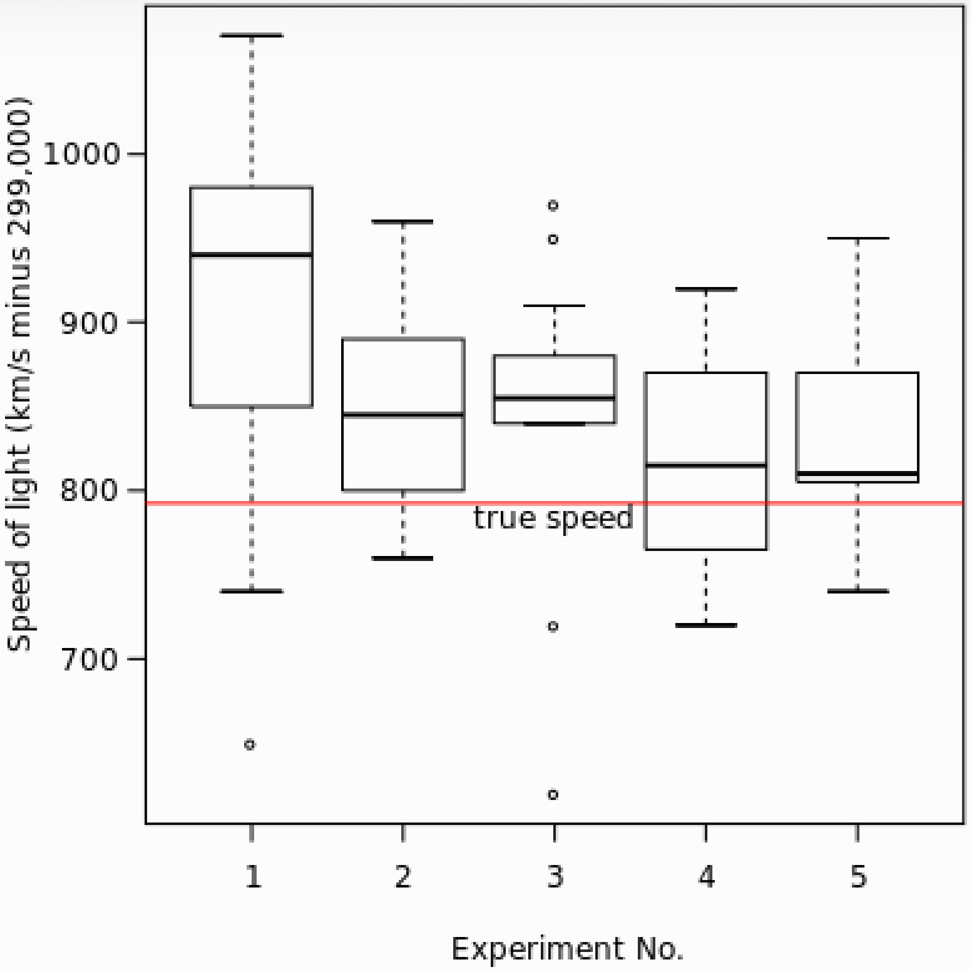
Normal Probability plots



Outliers detection

Box plot:

- Median
- Min value
- Max value
- Min Q
- Max Q
- Outliers



Outliers detection

Model-based are totally based on our model. For example, we could assume, that data are from normal distribution, then we could apply methods for that:

- Chauvenet's criterion
- Grubbs test for outliers
- Dixon's Q test
- etc.

Outliers detection

But in ML we usually do not have such assumptions, so anomaly detection techniques are applied.

Lecture plan

- Introduction into outliers
- Outliers types
- Outliers detection
- Anomaly detection
- Working with outliers

Anomaly detection

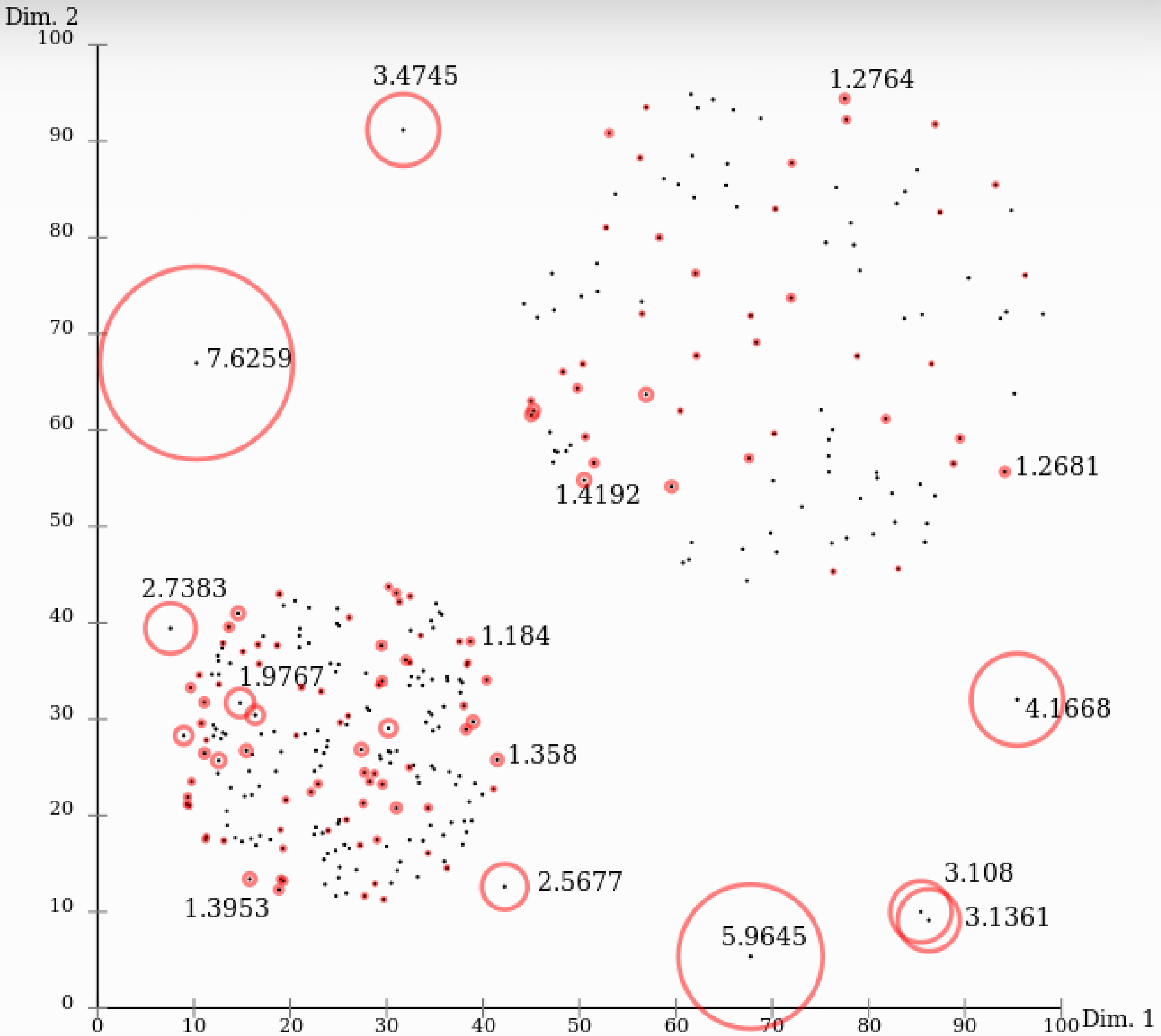
- Unsupervised – unlabeled data set, assume that majority of the instances in the data are normal
- Supervised – data set is labeled as “normal” and “abnormal”
- Semi-supervised – construct a model on a fully normal data set, and then for each new instance estimate its probability to be from this model

Methods for anomaly detection

- Density-based
- Subspace and correlation based
- One class SVM
- Replicator Neural Networks
- Cluster analysis based
- Deviations from association rules
- Fuzzy logic
- Feature bagging, score normalization

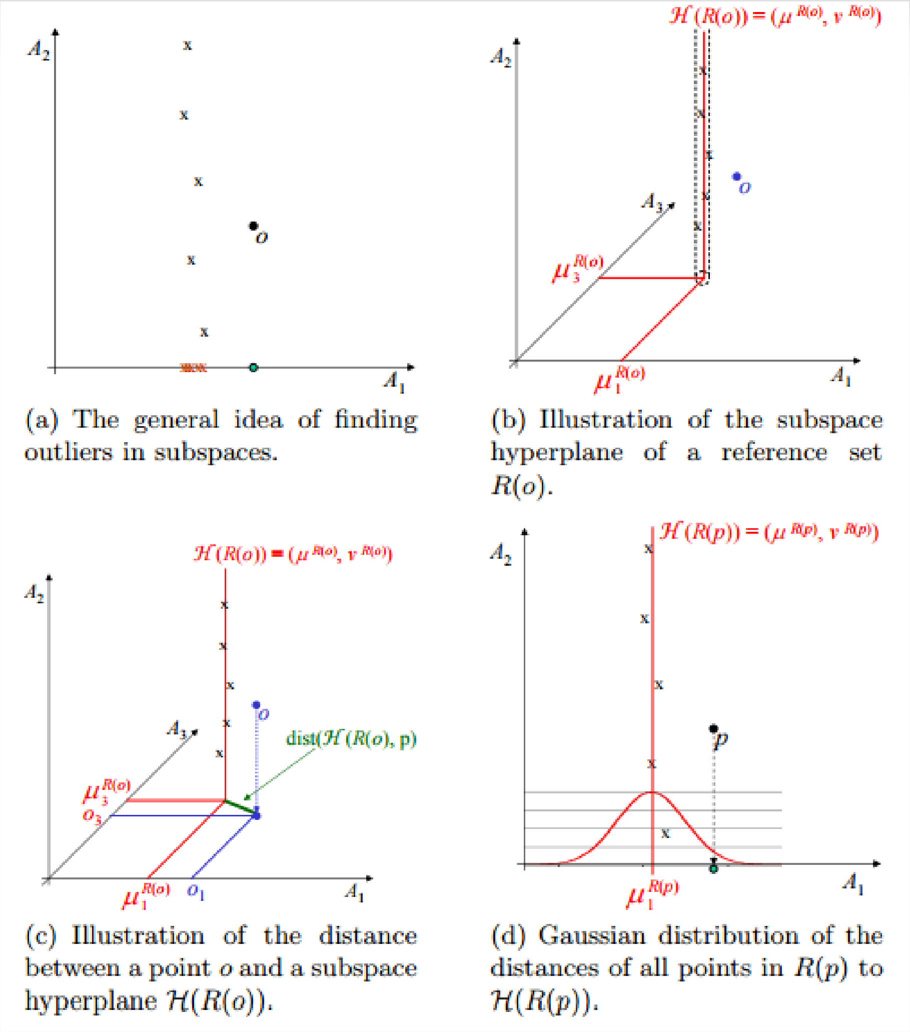
Destiny based

- Knn,
- Local Outlier Factor,
- etc.



Subspace and correlation based

Kriegel, H. P.; Kröger, P.; Schubert, E.; Zimek, A. (2009). *Outlier Detection in Axis-Parallel Subspaces of High Dimensional Data*. Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science. p. 831


A_2
A_1
o
$\mathcal{H}(R(o)) = (\mu^{R(o)}, \nu^{R(o)})$
A_3
$\mu_3^{R(o)}$
$\mu_1^{R(o)}$
o_3
o_1
dist($\mathcal{H}(R(o))$, p)
$\mathcal{H}(R(p)) = (\mu^{R(p)}, \nu^{R(p)})$
p
$\mu_1^{R(p)}$

(a) The general idea of finding outliers in subspaces.

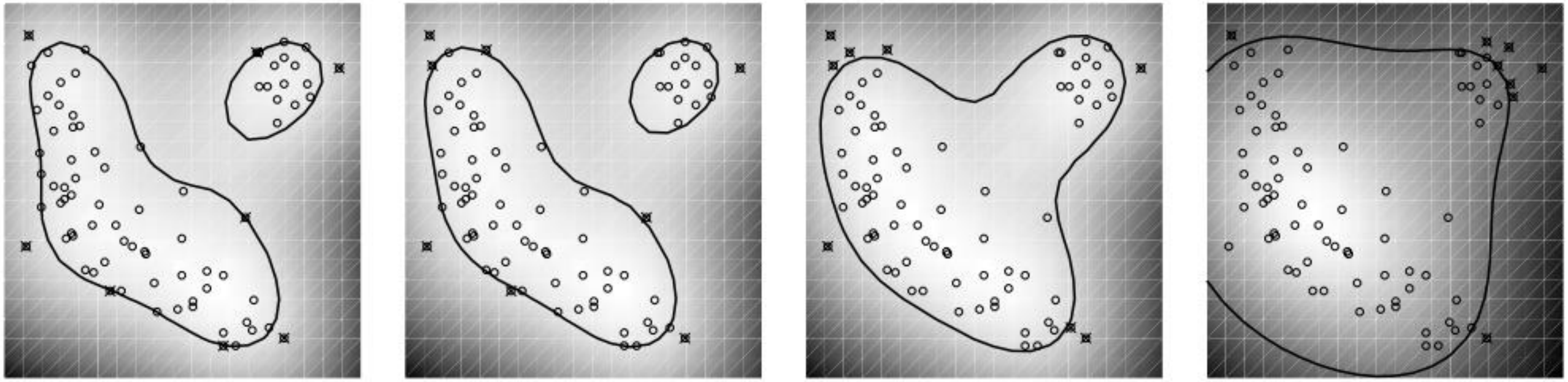
(b) Illustration of the subspace hyperplane of a reference set $R(o)$.

(c) Illustration of the distance between a point o and a subspace hyperplane $\mathcal{H}(R(o))$.

(d) Gaussian distribution of the distances of all points in $R(p)$ to $\mathcal{H}(R(p))$.

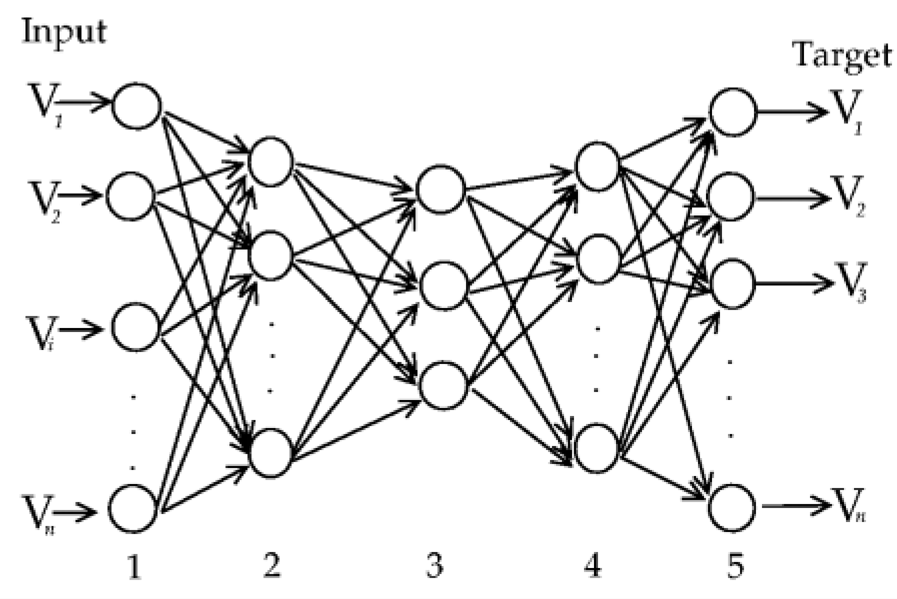
One class SVM

- Novelty detection method. Also applicable for outlier detection
- Training set should not be contaminated by outliers as it may fit them
- Ability to capture the shape of the data set
- Performing better when the data is strongly non-Gaussian, i.e. with two well-separated clusters



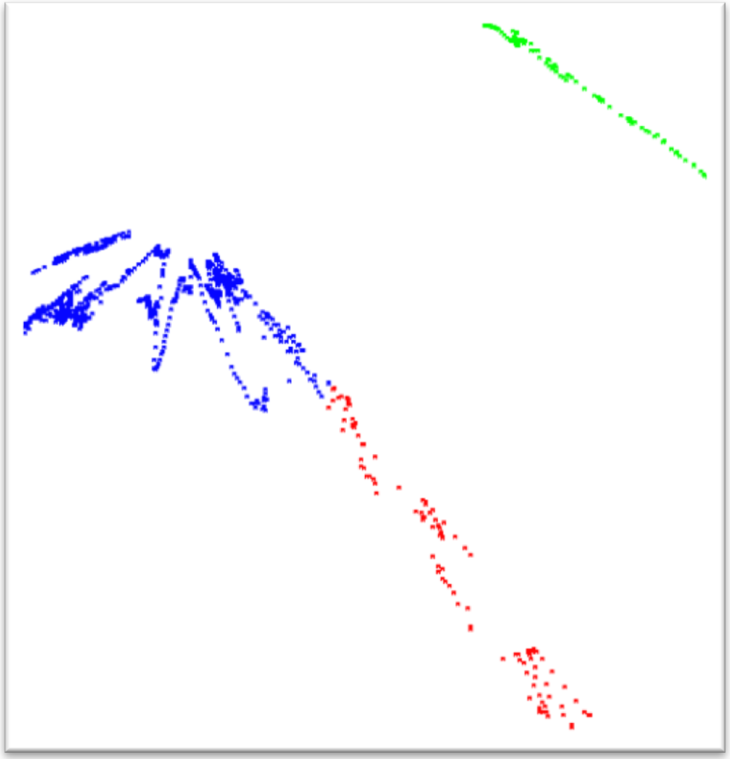
Replicator Neural Networks

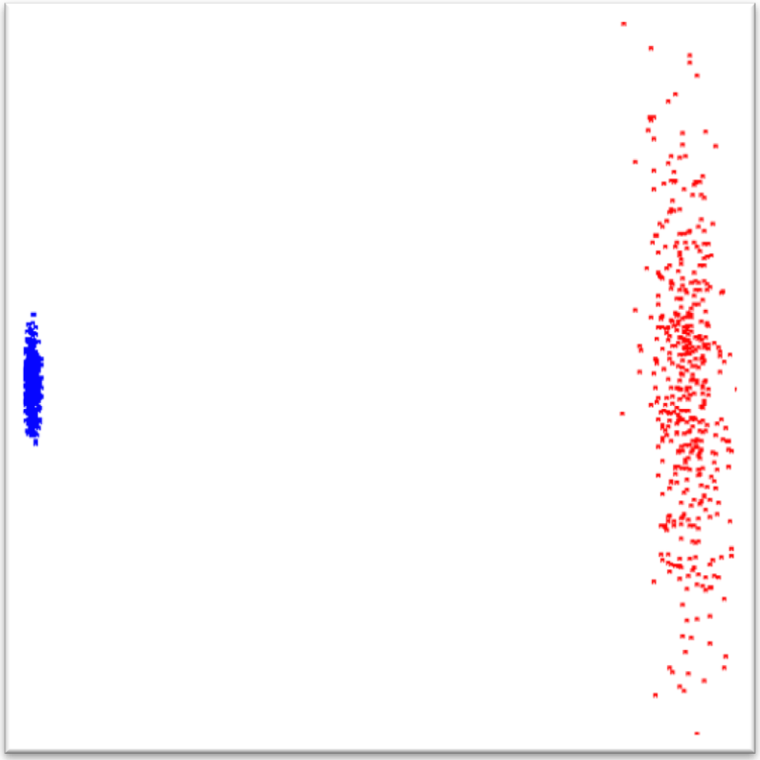
- Three hidden layers
- Number of inputs = number of outputs
- Network reconstructs some objects poorly = this objects are outliers



Cluster analysis based

Depending on the point of view different clusters may be considered as outliers



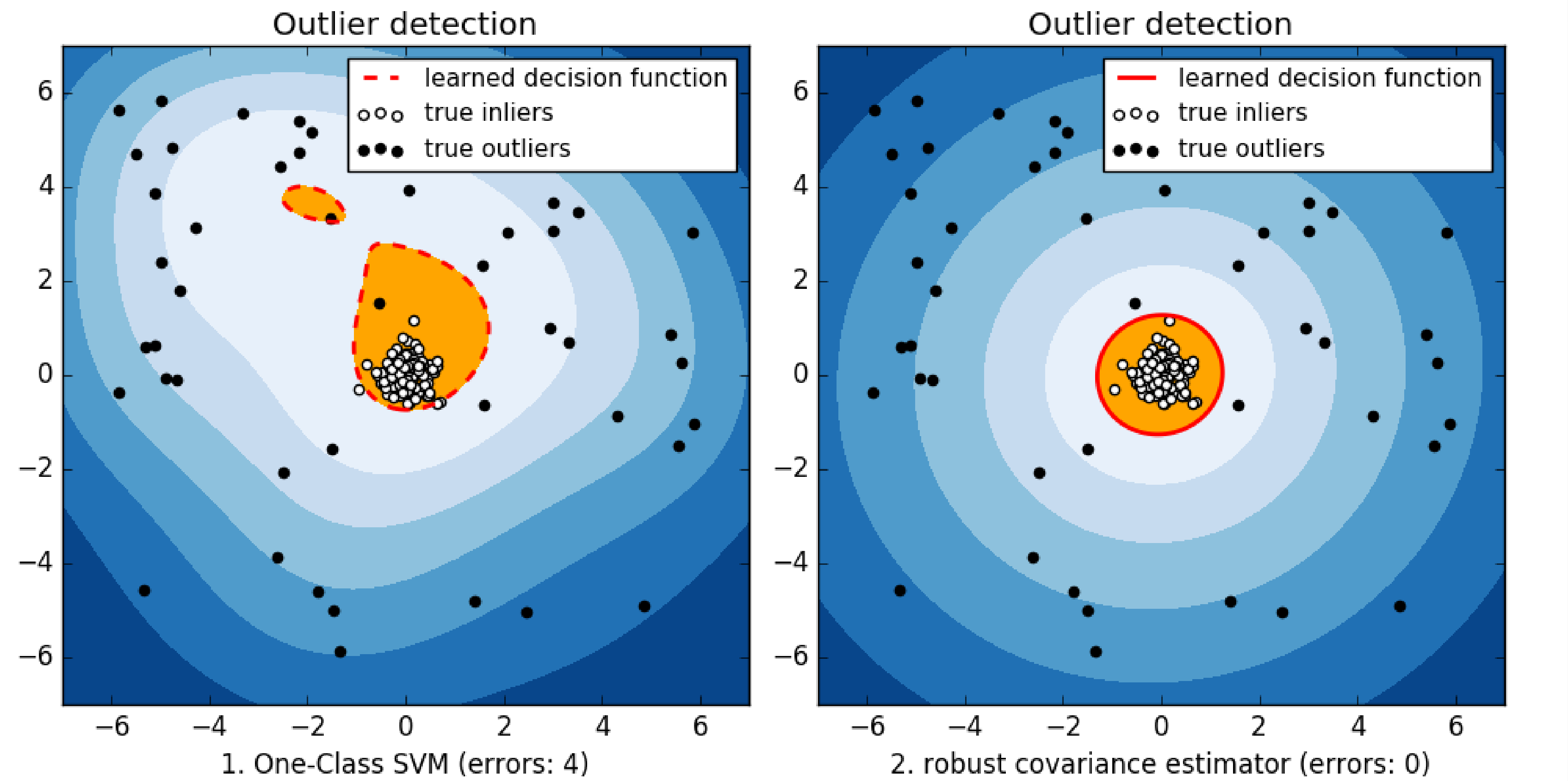


Methods for anomaly detection

- Density-based
- Subspace and correlation based
- One class SVM
- Replicator Neural Networks
- Cluster analysis based
- Deviations from association rules
- Fuzzy logic
- Feature bagging, score normalization

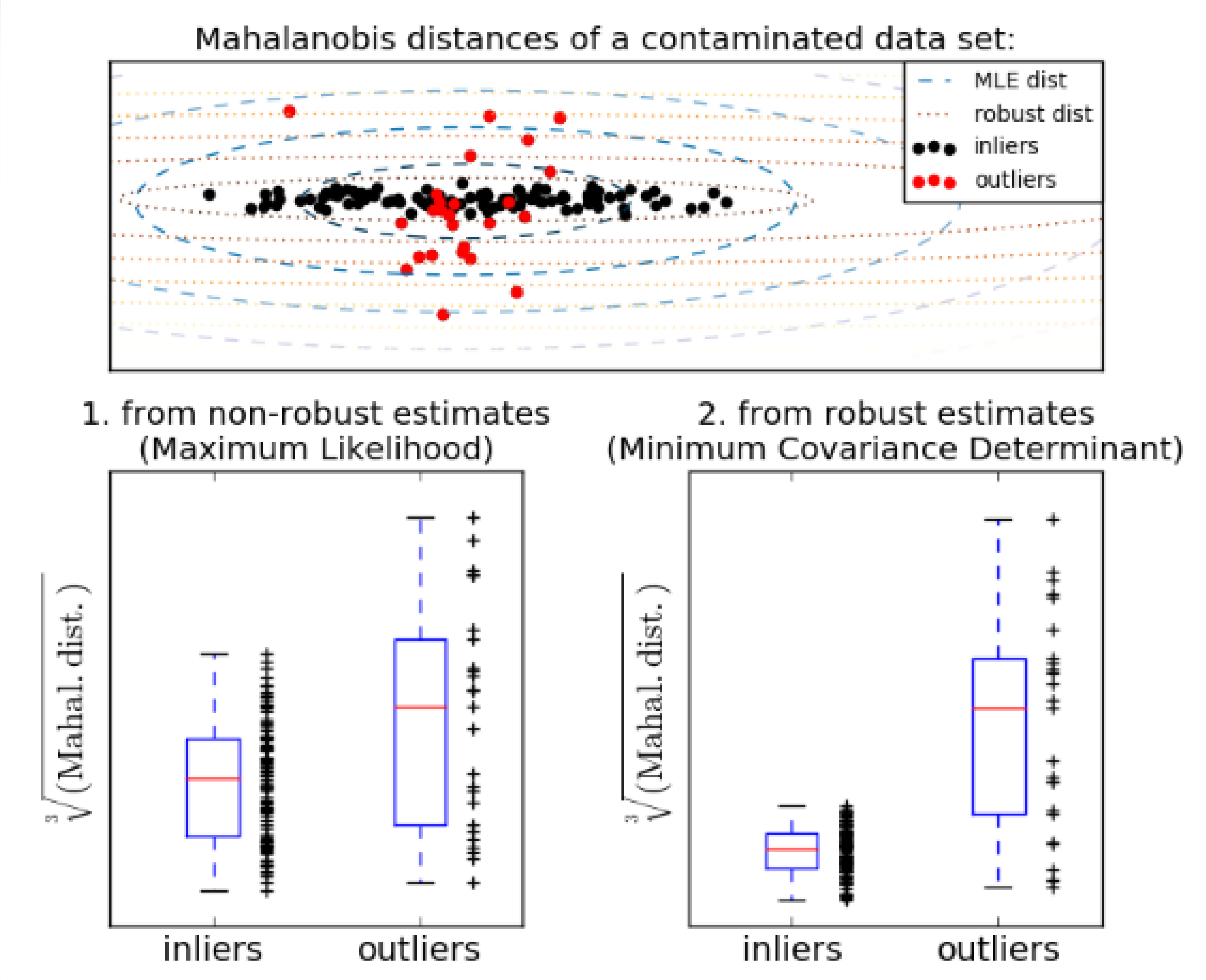
Pyhon

SVM: svm.OneClassSMV



Pyhon

covariance.
EllipticEnvelope



Weka

InterquartileRange

Makes new attributes, that are basically labels for Outliers and Extreme Values

Outliers:
$Q3 + OF*IQR < x <= Q3 + EVF*IQR$ or
$Q1 - EVF*IQR <= x < Q1 - OF*IQR$

Extreme values:
$x > Q3 + EVF*IQR$ or
$x < Q1 - EVF*IQR$

Lecture plan

- Introduction into outliers
- Outliers types
- Outliers detection
- Anomaly detection
- Working with outliers

Working with outliers

Retention vs Exclusion. Example – normal distribution is quite controversial.

Truncation vs Winsorising. Example - removal and replacement in time series.

Working with outliers

Some model examples:

- In regression model only points with high influence on the coefficients may be excluded.
- “Fat-tails” distribution – increased amount of extreme values.