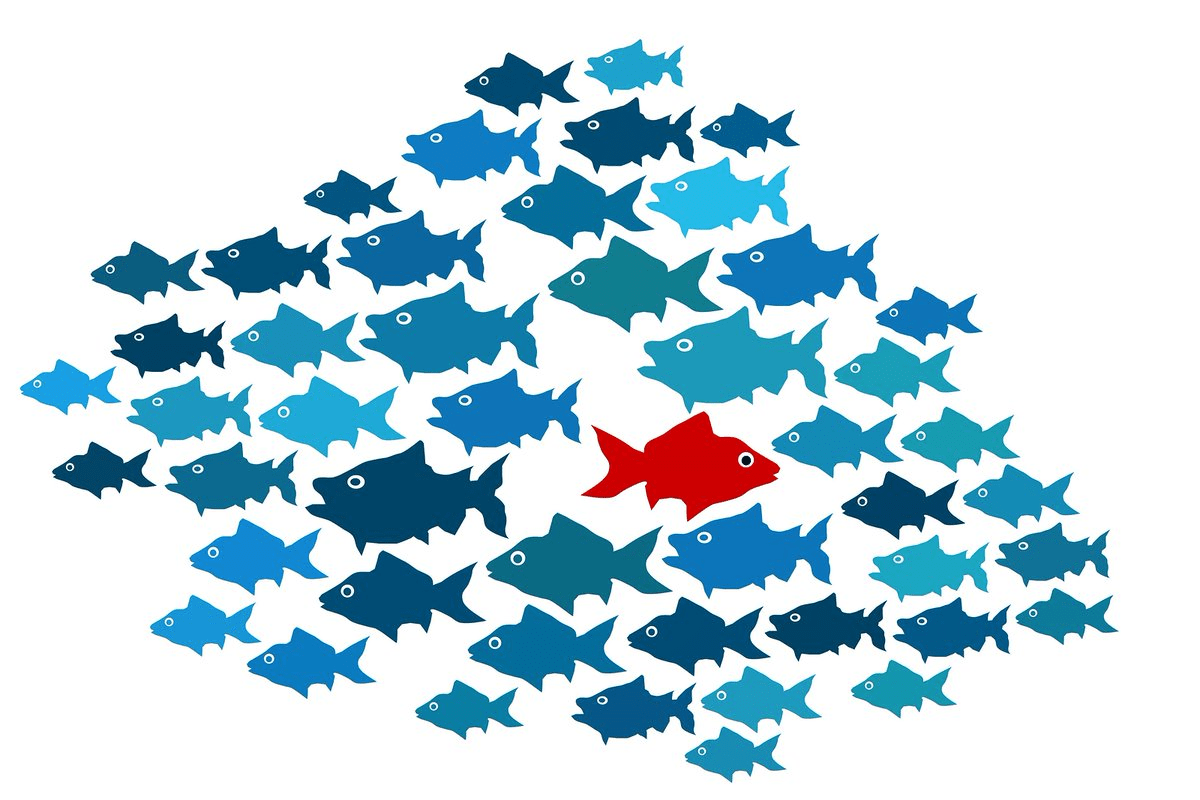
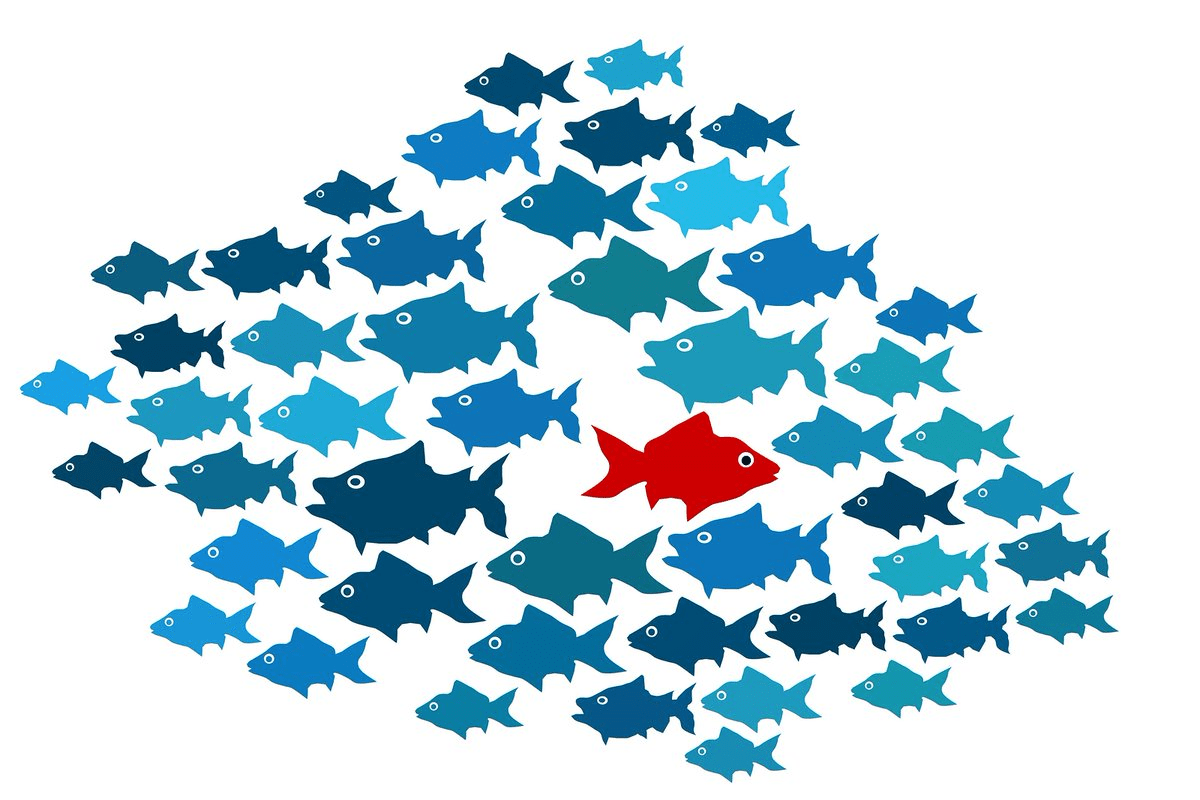
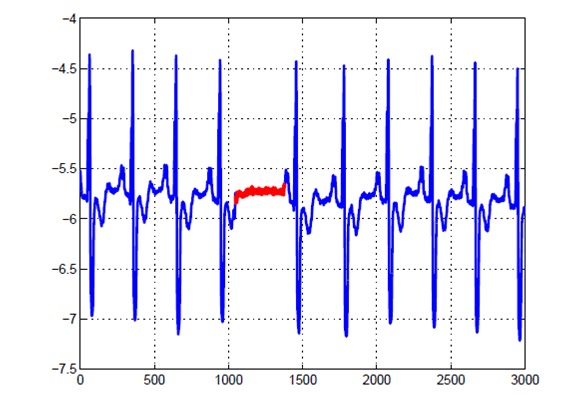
# Scratch book unsupervised learning 12

# Anomaly detection



Anomaly – anything that does not look like everything else.

* Point anomaly – one point that differs from every one.
* **Contextual anomaly**  - a point that differs from is environment.
* Collective anomaly



# Unsupervised anomaly detection

* Statistical - We build a model for the data with the anomalies.
* Proximity-based – An anomalous point is far from other points.
* Density-based - An anomalous point has a low density.
* Clustering-based - An anomalous point does not sit well in any cluster.

# Statistical anomaly detection

1. Define a statistical model on the data
2. Find improbable data points.

Example I – 1D gaussian data. - Anomalous points are points that are K SD from average.

Example II – Define GMM on the data and compute points with low probability in GMM.

1. Throw a coin if face then anomalous, if number then not. – probably lambda to be anomalous.
2. Each category has a different distribution.

Fit the data to the model.

More formally:

Assume a latent variable that defines for each point – with probability lambda anomalous, with probability (1-lambda) I am normal. Normal points have a distribution P1. We assume known \lambda and P.

=

1. Assume that I know, which point is anomalous (A) and which point is not (N) and I have |A| number of anomalous point, |N| number of normal points.

=

=

Derivate the last term by theta and obtain the ML parameters of the model.

1. Assume, I know the parameters theta, I can decide if a point is anomalous if its probability in the normal model is bellow some threshold.
2. Return to 1, until convergence.

# Proximity-based outlier detection



Outlier score is distance to kth nearest neighbor.

Problem – sensitive to the selection of k (k=5)

Also, cannot detect contextual anomalies.



# Density-based outlier detection

An anomaly is not dense (compared to its neighbors)

1. DBSCAN
2. K Density
   * Sensitive (slightly less than distance based) to k
   * Expensive to compute if not implemented (Can cost up to if not using a KD-tree, else ).

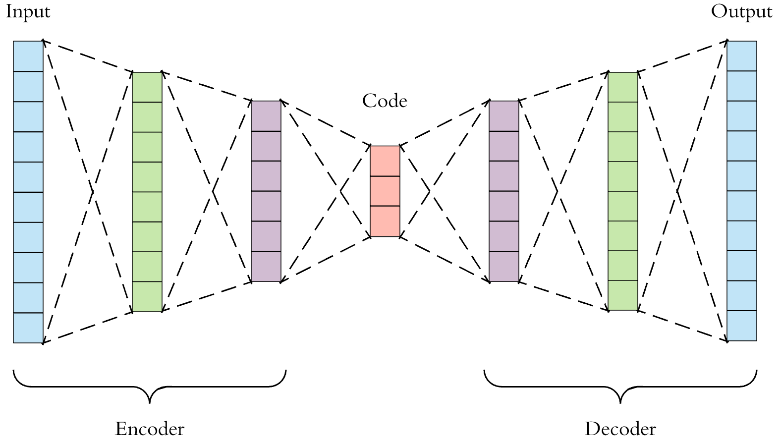
# Cluster-based outlier detection

1. Cluster all the data (usually GMM/FCM/K means).
2. Remove all the points that have a low clustering score (negative silhouette score, or likelihood in GMM).
3. Go back to 1 until we find no points to remove.

# Machine learning based anomaly detection

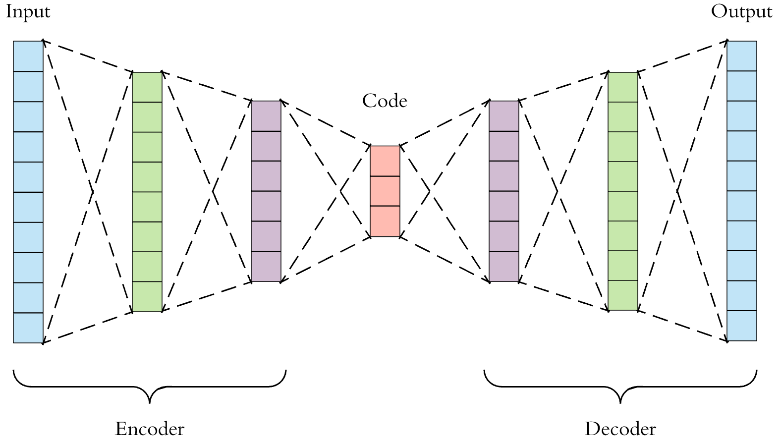
* One class classifier.
* Autoencoder.

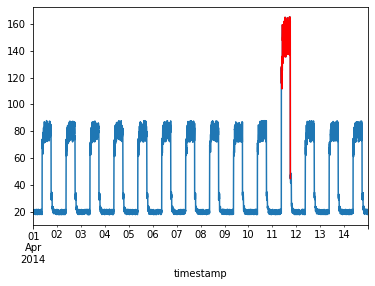
# Autoencoder



Train a machine to reproduce precisely the input by passing through a narrow layer.

x(t)

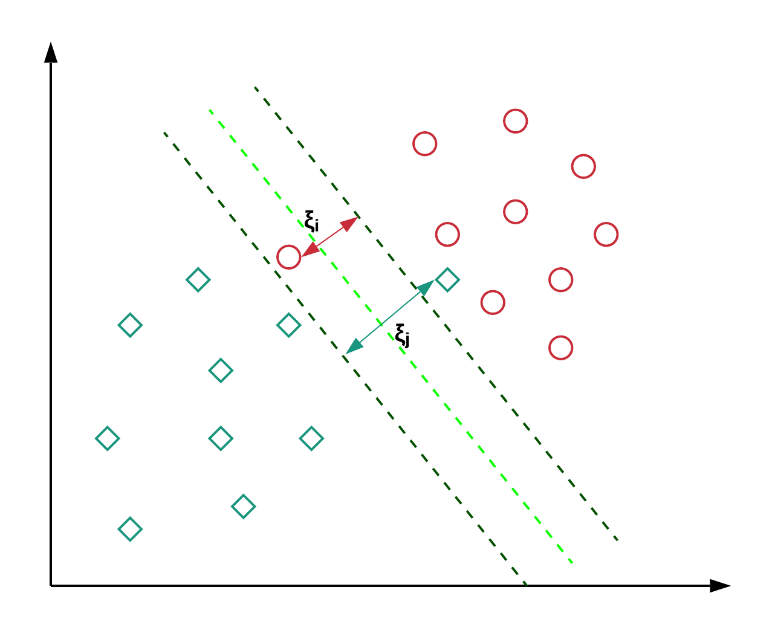
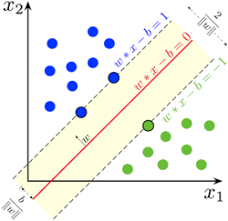




# One class SVM

SVM – support vector machine

Given input (a vector) and tag . I want a linear function s.t if , and if .



# One class SVM.