Anomaly Detection

Qinliang Su （苏勤亮）

Sun Yat-sen University

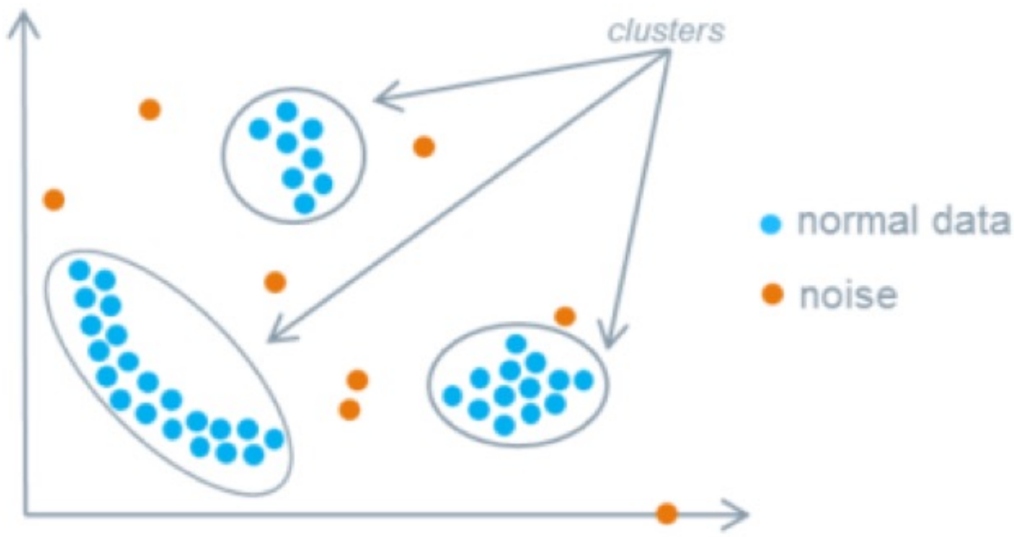
suqliang@mail.sysu.edu.cn

Outline

- Introduction
- Distance-Based Approach
- Density-Based Detection Approach
- Reconstruction-Based Detection Approach
- One-Class Classifier Approach
- Evaluation Metrics

What is Anomaly Detection?

- Goal: Detect or discover data samples that *deviate significantly from the majority of data samples*



- Anomaly detection sometimes is also called 'outlier detection'（离群点检测）or 'novelty detection'（新颖性检测）

What’s special in Anomaly Detection Tasks

- Normal data
  - Normal data can be collected easily and cheaply
  - Thus, normal samples are often assumed to be abundant

- Anomalous data
  - Anomalies are very diverse, and generally cannot be characterized explicitly
  - Exhaustive enumeration of anomalies is impossible
  - Informally, any samples that look significantly different from the normal ones can be viewed as anomalies

Typical Applications of Anomaly Detection

1) Network intrusion detection

Detecting unauthorized intrusion by monitoring the events occurring in computer or networks

- Challenges of traditional intrusion detection systems based on the signatures of known attacks
  - Can only be used to detect known attacks
  - However, sophisticated attacking methods are emerging every day

- By viewing it as an anomaly detection problem, any event that looks different from normal cases will be identified
  - Addressing the limitations of traditional methods

2) Fraud detection

Detecting criminal activities occurring in commercial organization

- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobil/cell Phone fraud
  - Insider trading in stock market
- Challenges

New fraud methods emerge from time to time. Traditional fraud detection methods fail to recognize the novel ones

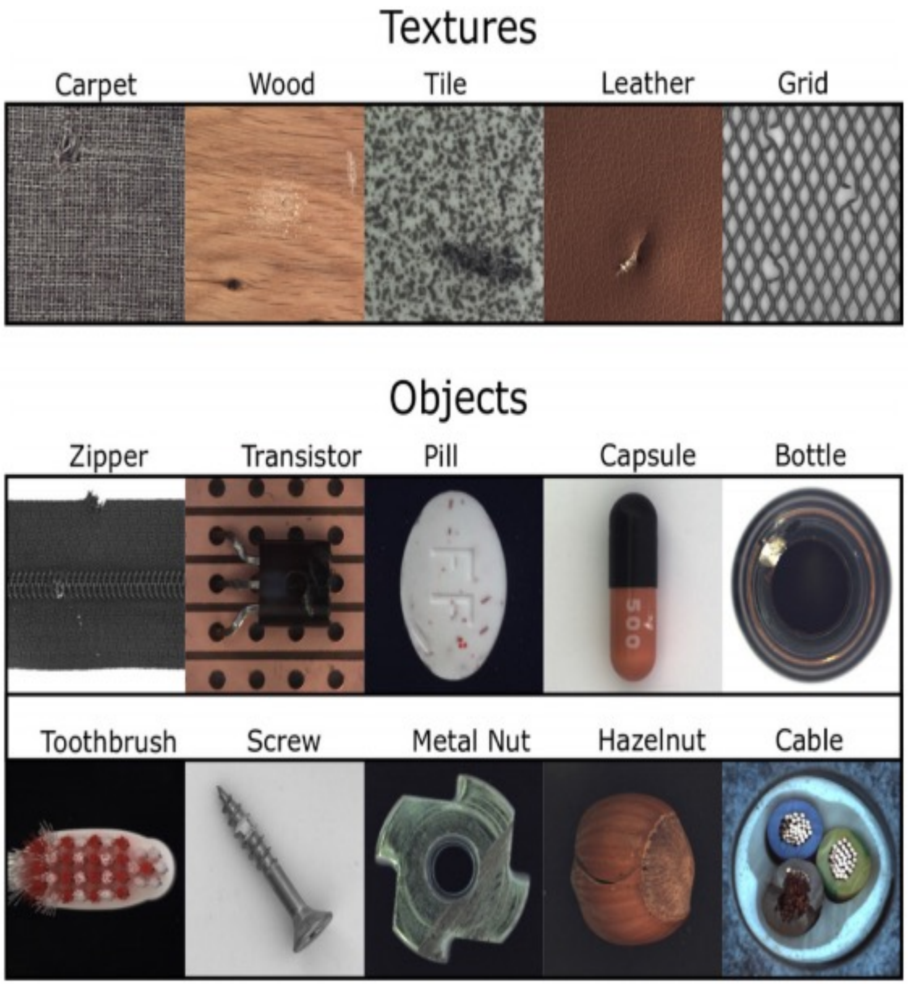
3) Rare disease detection

- Rare diseases
  - Diseases that rarely occur, *e.g.*, some types of cancers

- Specials in the task
  - Normal records are abundant
  - Types of rare diseases are inexhaustible, and the data of some types may be very scarce or even does not exist

- If diagnosis is carried out by examining the similarity between a patient's CT or X-ray and those of existing diseases, the rare diseases may not be recognized timely

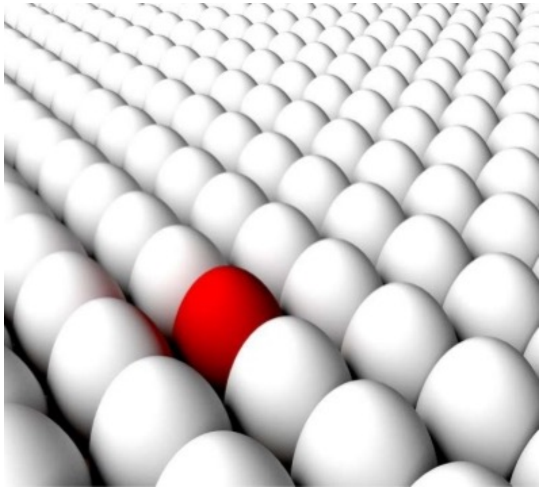
4) Industrial damage detection

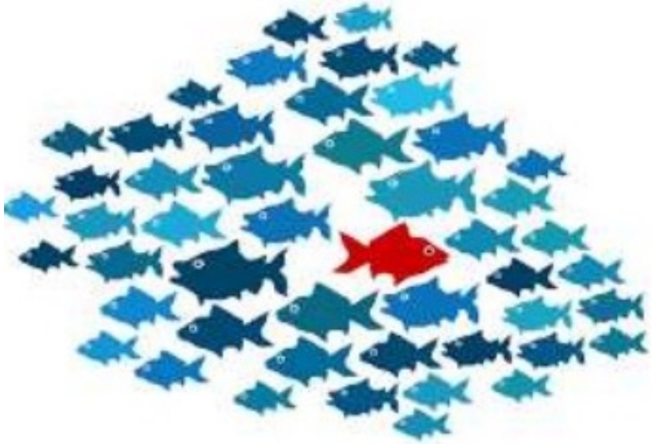


- Flawed manufacturing
  - Texture is damaged
  - Object surface is broken
  - Object inside is fractured
- Specials in the task
  - Damage types could vary widely
  - Unable to describe every type of damage accurately
  - The normal samples are abundant

Anomaly Detection Perspective

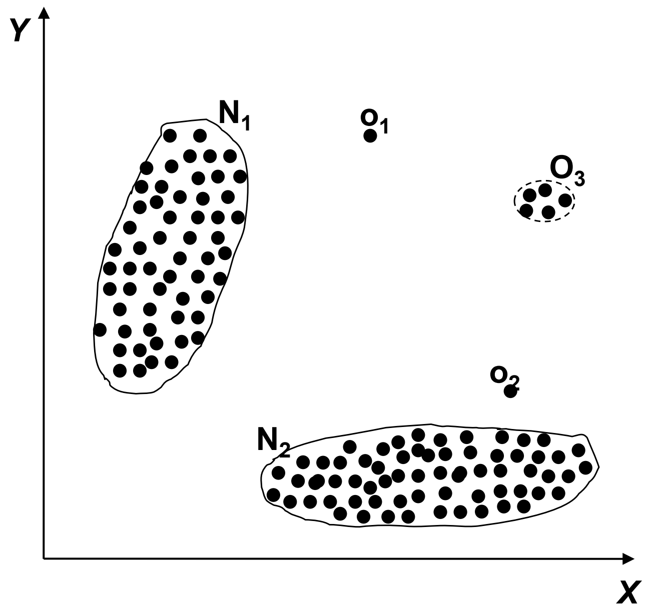
- All of the aforementioned applications share the characteristics:
  - The ‘data of interest’ varies in forms dramatically. Impossible to give them an exact description.
  - Ordinary data is abundant
- By viewing the ‘data of interest’ as anomalies and ordinary data as normal, these tasks can be considered as anomaly detection





Types of Anomaly

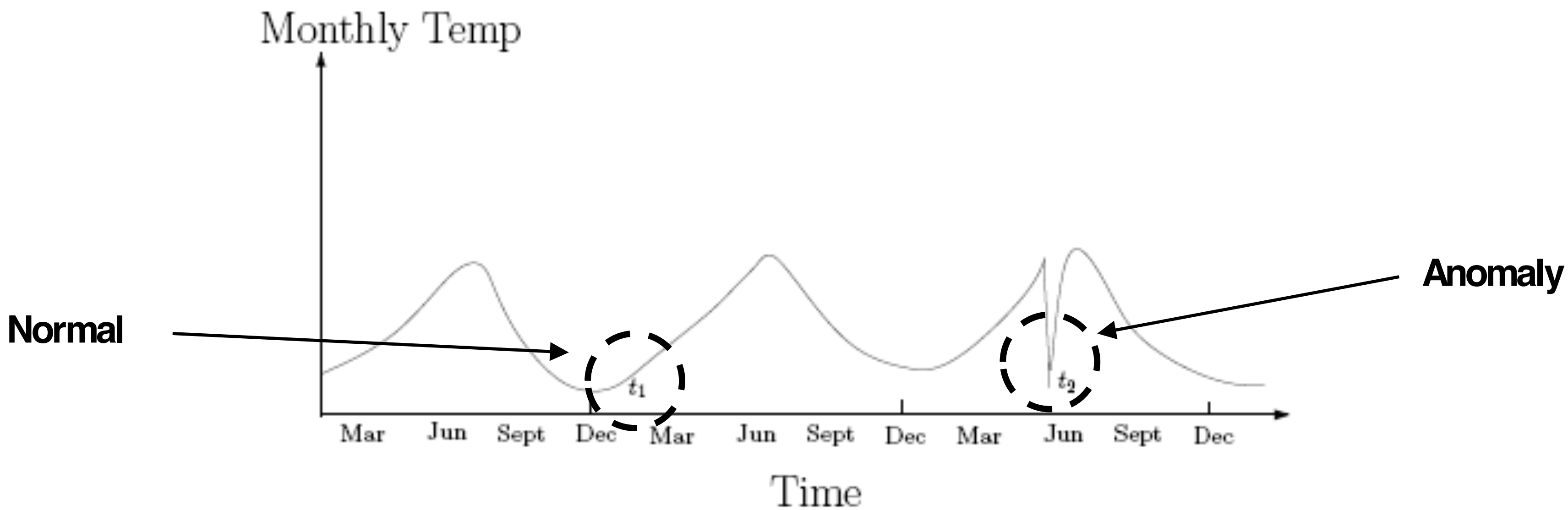
- Point anomaly



  - Individual data point can be determined as an anomaly or not *by itself*

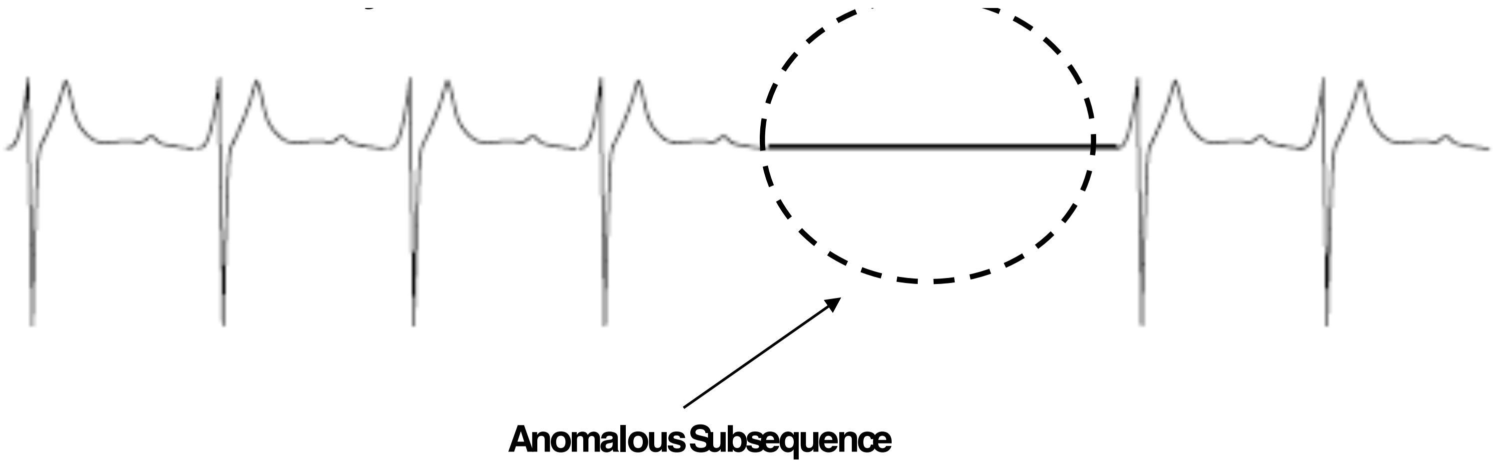
Point anomaly detection is the focus of this lecture

- Contextual anomaly



  – Individual instances look normal
  – But when they are examined within a context, they may look anomalous

- Collective anomaly



  - Individual instances look normal
  - When a collection of instances are examined together, they may be deemed as anomalous

Outline

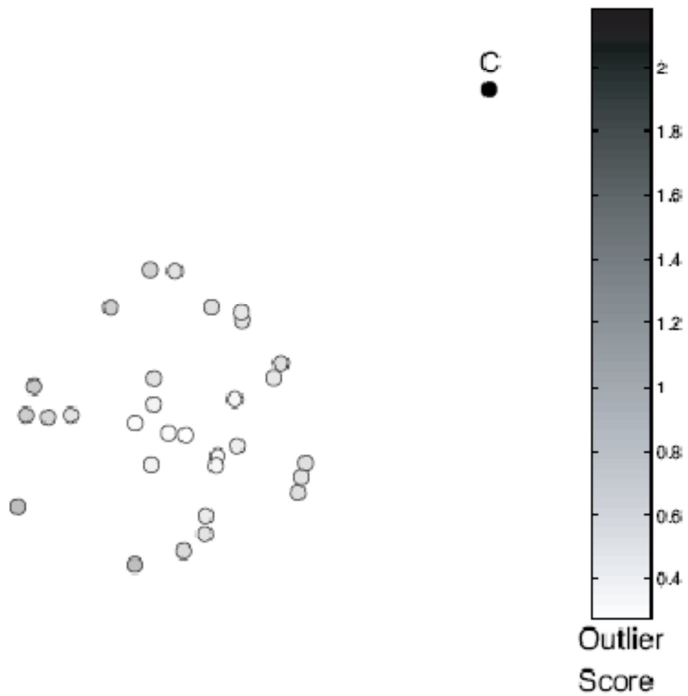
- Introduction
- Distance-Based Approach
- Density-Based Detection Approach
- Reconstruction-Based Detection Approach
- One-Class Classifier Approach
- Evaluation Metrics

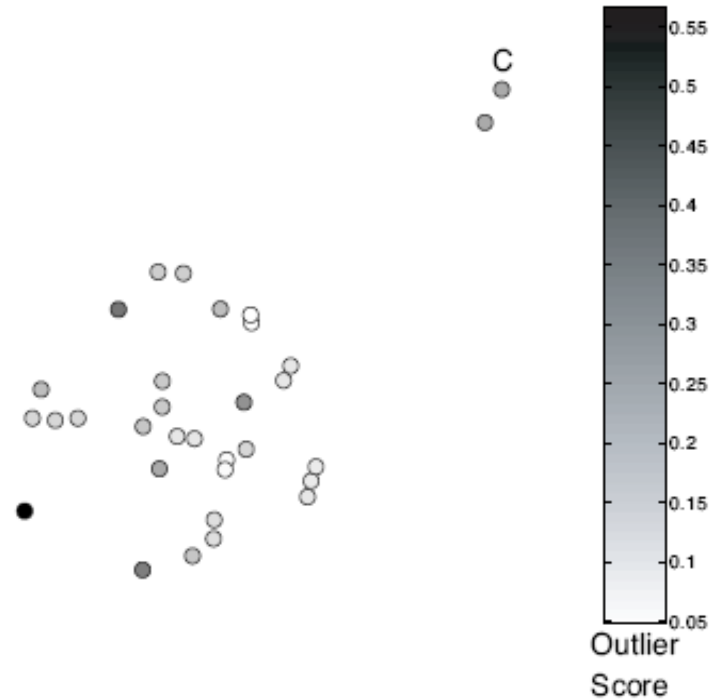
- Idea behind the distance-based approach

Anomalies are the instances that are far away from the majority

- Common method
  - Computing the outlier score as the distance to the k-th nearest neighbor
  - Using the score to determine whether an instance is anomalous or not

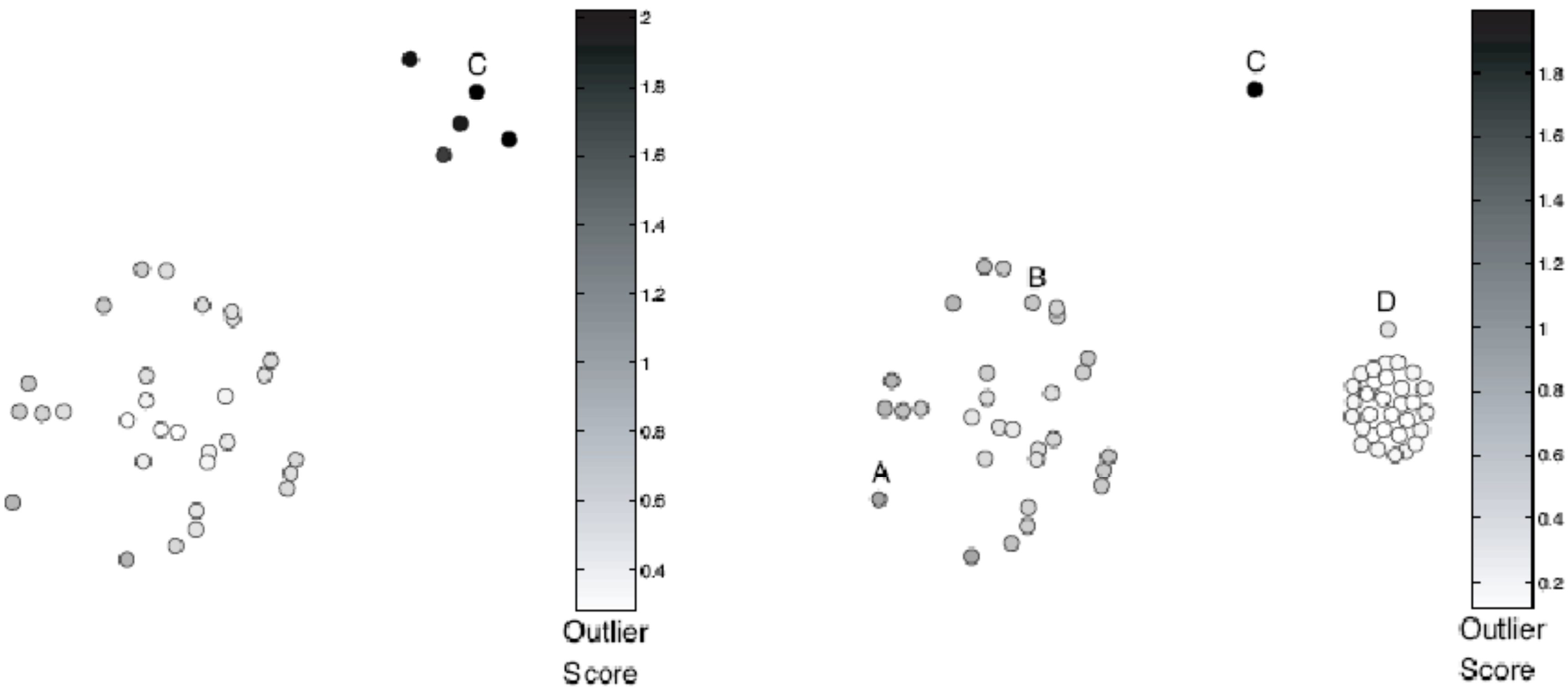
- Obviously, the method is sensitive to the value k





$Outlier\ score =$Distance to the 5-th neighbor

$Outlier\ score =$Distance to the 1-th neighbor



$Outlier\ score =$Distance to the 5-th neighbor

- Pros
  - Intuitive and easy to understand
  - Interpretable

- Cons
  - Complexity is high $O(n^2)$
  - Sensitive to the value k
  - Difficult to find a good distance measure, especially for high-dimensional data, *e.g.*, images

Outline

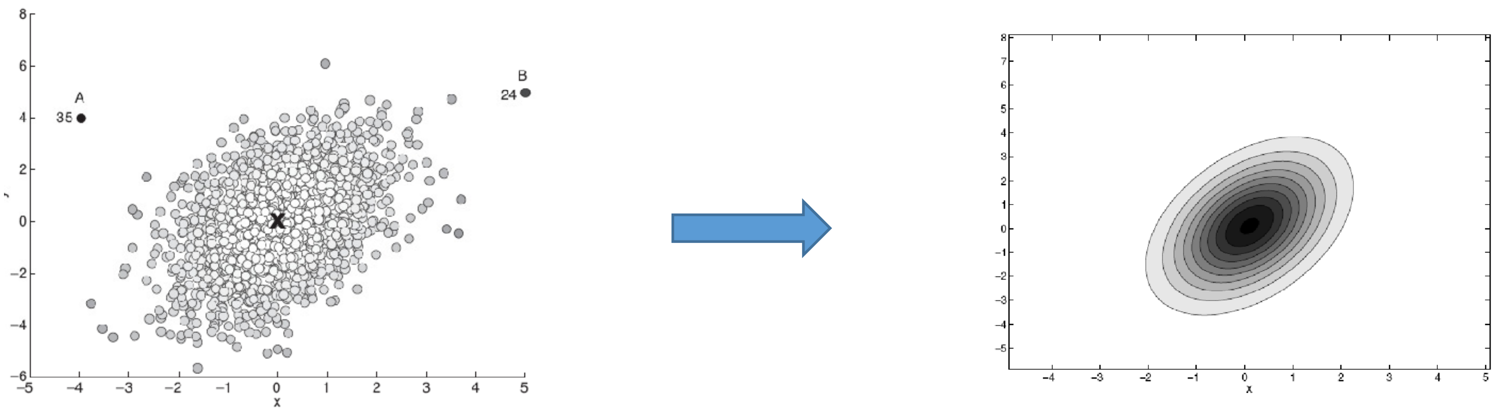
- Introduction
- Distance-Based Approach
- Density-Based Detection Approach
- Reconstruction-Based Detection Approach
- One-Class Classifier Approach
- Evaluation Metrics

- Idea behind the density-based approach

Anomalies are the instances that fall in the low-density region

- Procedures of this approach

1) Estimate the probability density distribution of normal data from a given set of normal data instances, *e.g.*, $\hat{p}(\boldsymbol{x})$



2) For a new instance $\boldsymbol{x}_{new}$, if its density $\hat{p}(\boldsymbol{x}_{new})$ is smaller than a threshold, we deem $\boldsymbol{x}_{new}$ is an anomaly

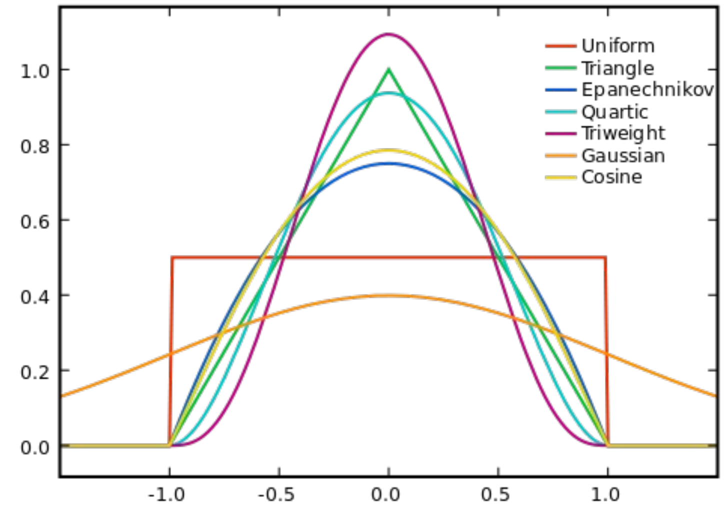
How to Estimate the Density?

1) Kernel density estimation (KDE)

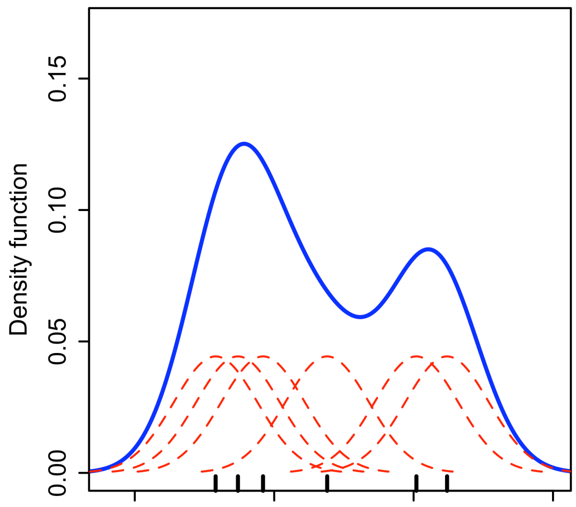
- Given a set of normal data instance $\{x_i\}_{i=1}^{N}$, the probability density function $p(x)$ can be estimated as

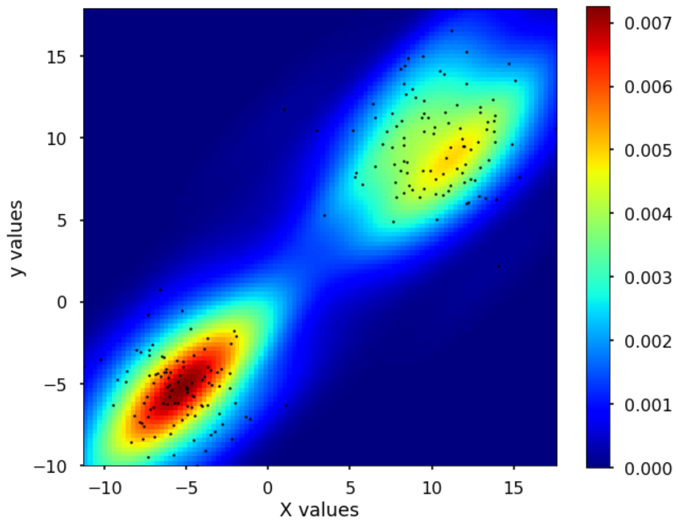
$$p(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right)$$

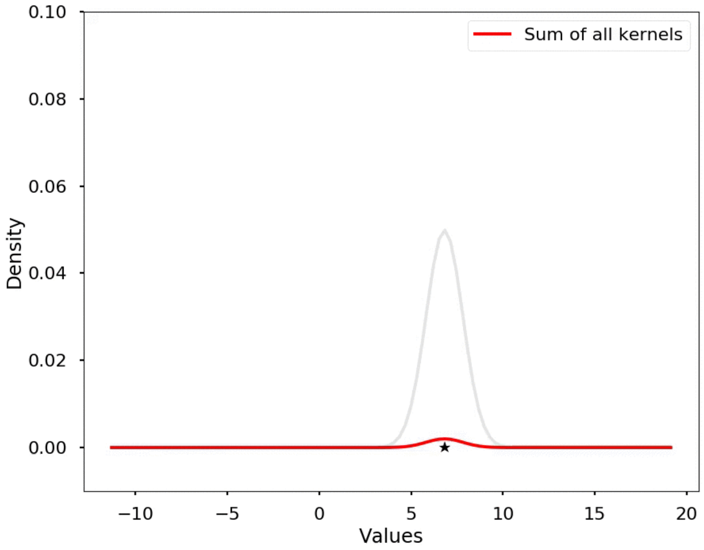
  - $K(\cdot)$ is a kernel function, *e.g.*, Gaussian kernel, uniform kernel, triangle kernel etc.
  - h is a parameter controlling the smoothness (bandwidth)



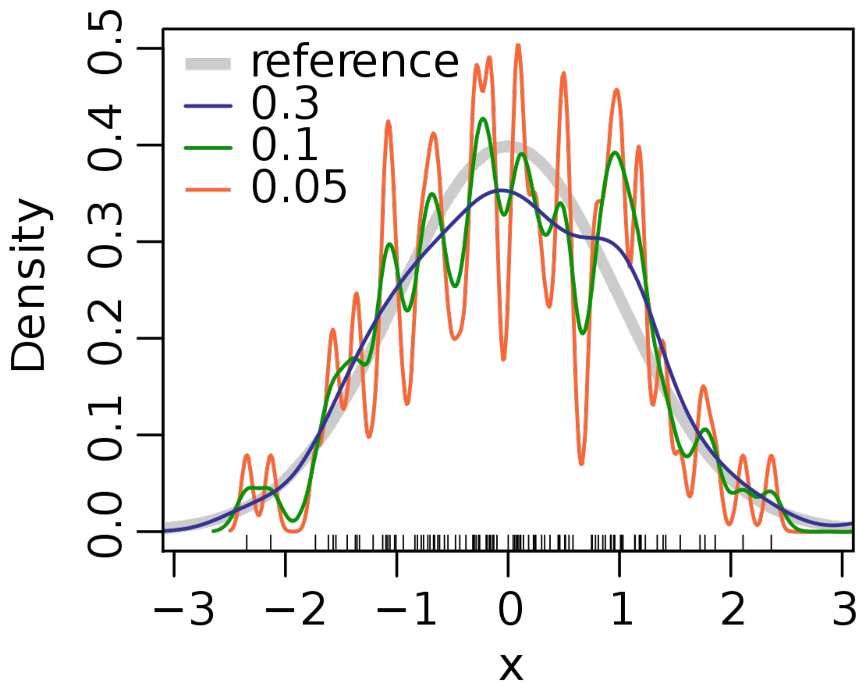
- Examples


Density function
0.00
0.05
0.10
0.15


y values
X values
15
10
5
0
−5
−10
−10
−5
0
5
10
15
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000


Sum of all kernels
Density
Values
0.10
0.08
0.06
0.04
0.02
0.00
−10
−5
0
5
10
15
20

- Limitations of KDE
  - Very sensitive to the parameter h



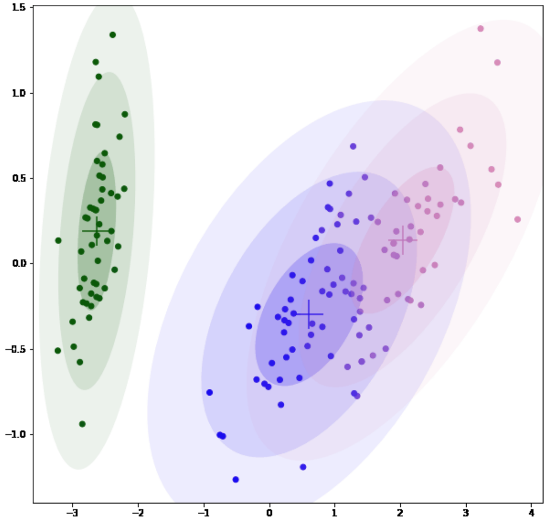
  - Struggling in high-dimensional scenarios, *e.g.*, images

2) Fitting the data with a known density function

- Given a collection of normal instances $\{\boldsymbol{x}_i\}_{\{i=1\}}^{N}$, we train a known distribution (*e.g.* Gaussian, Gaussian mixture) to fit them

$$\boldsymbol{\theta} = \arg\max_{\boldsymbol{\theta}} \frac{1}{N}\sum_{i=1}^{N} \ln p(\boldsymbol{x}_i; \boldsymbol{\theta})$$

  - where $p(\boldsymbol{x};\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ or $p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$



- Limitations
  - The representational ability of known distributions are limited
  - Struggling in modeling high-dimensional data, *e.g.* images

3) Two-stage: dimension reduction + density learning

The first two methods perform poorly on high-dimensional data

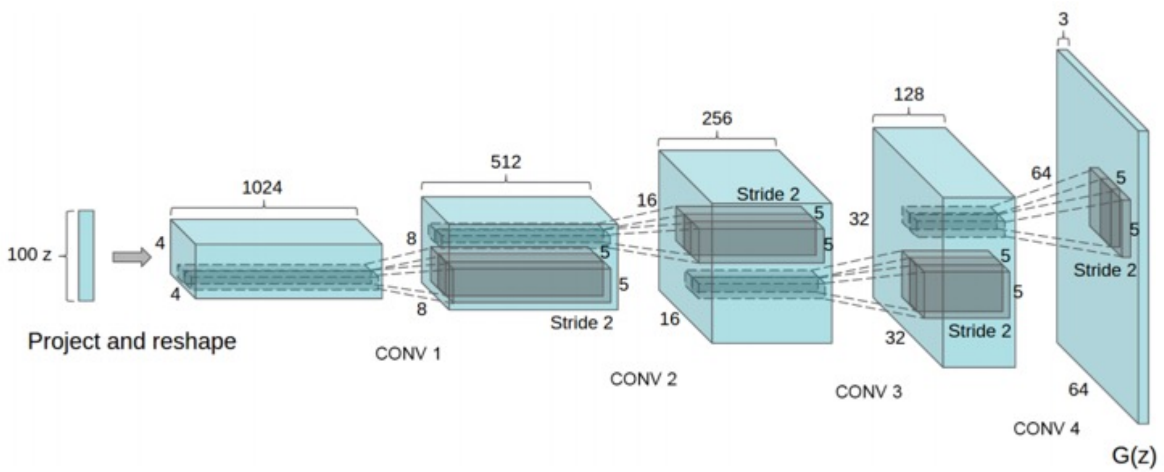
- Approaches to address this issue
  - A direct way is to first learn a low-dimensional representation for each data instance

    Remark: we may resort to PCA, auto-encoder or other methods to extract low-dimensional representations

  - Then, apply the KDE or density fitting methods on the low-dimensional representations

4) Deep Generative Models

- For complex data, we can train a deep generative model on normal data instances so that the probability distribution of normal data $p(\boldsymbol{x})$ can be learned approximately

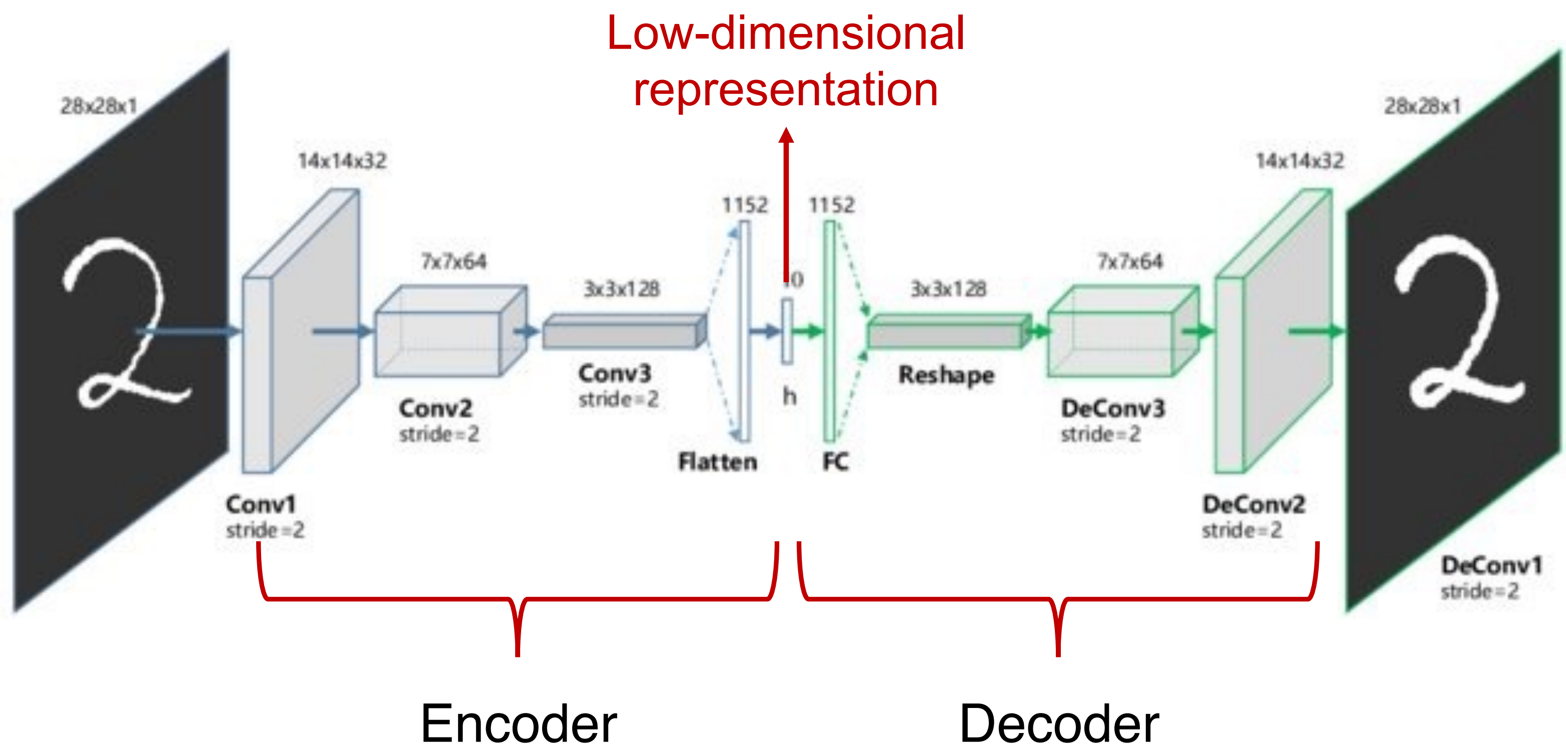


- There are many kinds of deep generative models, *e.g.*,
  – Variational auto-encoders (VAE)
  – Generative adversarial networks (GAN)
  – Energy-based models

Outline

- Introduction
- Distance-Based Approach
- Density-Based Detection Approach
- Reconstruction-Based Detection Approach
- One-Class Classifier Approach
- Evaluation Metrics

- Training an auto-encoder *on the normal data instances*
  - First compressing the high-dimensional data into a low-dimensional representation
  - Then, seeking to recover the original data with the compressed representation



- Since the auto-encoder is trained on normal instances, when a testing data is fed into the model, we may expect to observe the following phenomena

Normal data instances

- it can be reconstructed well if *the input data is normal*

Input

Reconstructed

- it cannot be reconstructed well if *the input data is an anomaly*

Input

Reconstructed

Outline

- Introduction
- Distance-Based Approach
- Density-Based Detection Approach
- Reconstruction-Based Detection Approach
- One-Class Classifier Approach
- Evaluation Metrics

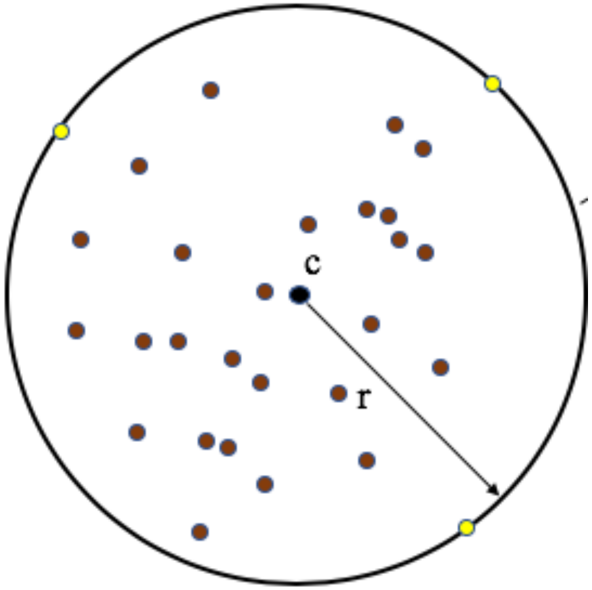
- Support vector data description (SVDD)

  Idea: finding the smallest hypersphere (with radius r and center c) that can encompass all of the data instances

$$\min_{r,c} \; r^2$$

$$s.t. \quad \|\Phi(x_i) - c\|^2 \le r^2$$

$$\forall \, i = 1, 2, \cdots, N$$



  – $\Phi(\cdot)$ could be any nonlinear mapping

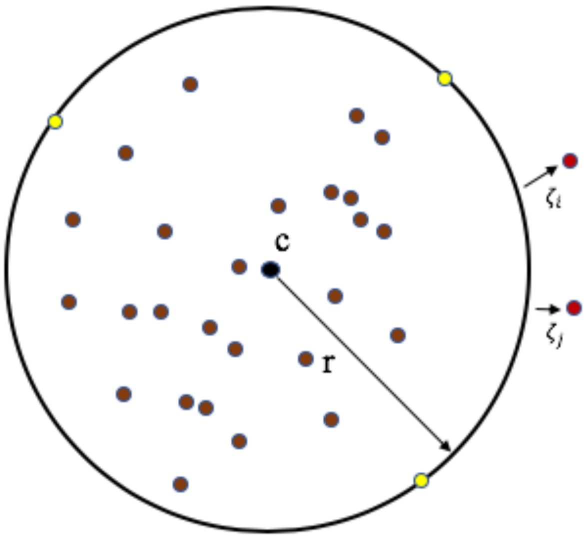
➢ However, this formulation is very sensitive to the presence of outliers

- To have the model more flexible, we modify the formulation as

$$\min_{r,c,\zeta_i} r^2 + \frac{1}{\nu n}\sum_{i=1}^{n}\zeta_i$$

$$s.t. \quad \|\Phi(x_i) - c\|^2 \leq r^2 + \zeta_i$$

$$\zeta_i \geq 0, \quad \forall\, i = 1, 2, \cdots, N$$



  - The slack variables ζ_i allow a soft-boundary
  - The hyper-parameter $\nu \in (0, 1]$ approximately controls the proportion of instances outside of the sphere

- With the optimal r^* and c^*, a testing instance is judged as an anomaly or not by *checking whether it locates in the sphere*

- The feature function $\Phi(\cdot)$ could be any nonlinear mapping

1) $\Phi(\cdot)$ could be the one derived from the kernel function $k(\cdot,\cdot) = \Phi^{\mathrm{T}}(\cdot)\Phi(\cdot)$

   – The optimization problem can be solved in its dual form

   – Giving rise to a SVM-like solution that is expressed in form of support vectors

2) $\Phi(\cdot)$ could be a deep neural network

   – The optimization problem can be solved by minimizing the following unconstrained loss function with SGD algorithms

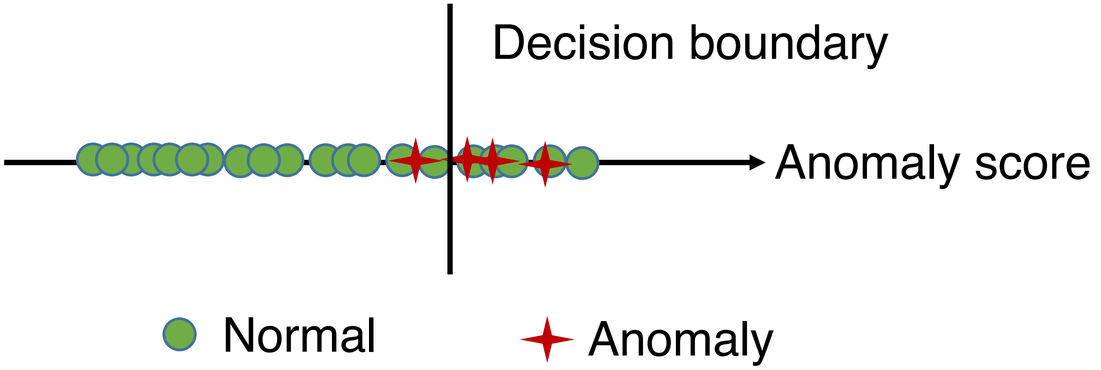
$$\mathcal{L} = r^2 + \frac{1}{\nu n}\sum_{i=1}^{n} \max(\|\Phi(x_i) - c\|^2 - r^2, 0)$$

Outline

- Introduction
- Distance-Based Approach
- Density-Based Detection Approach
- Reconstruction-Based Detection Approach
- One-Class Classifier Approach
- Evaluation Metrics

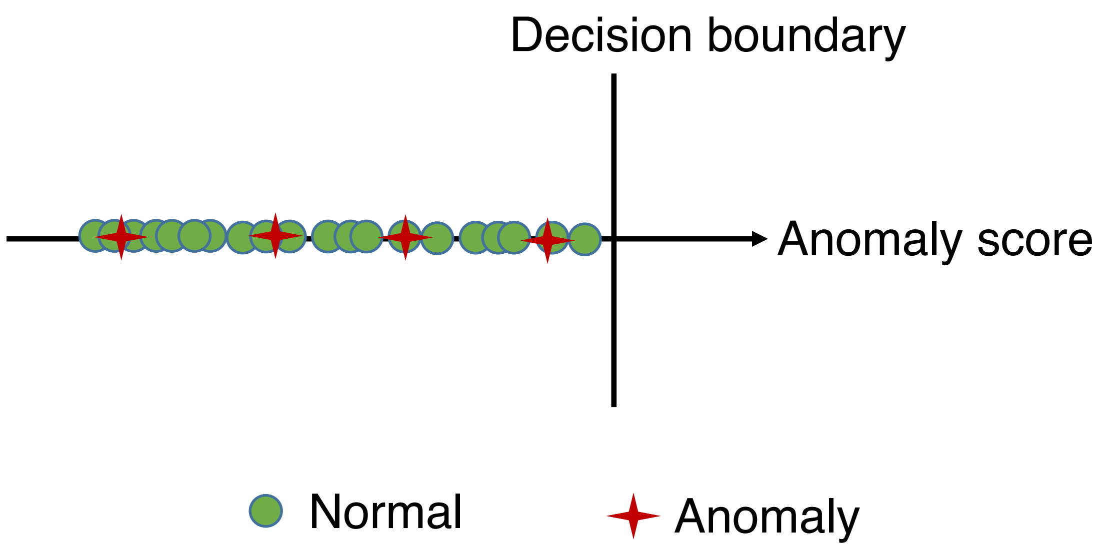
ROC Curve

- Possible detection results



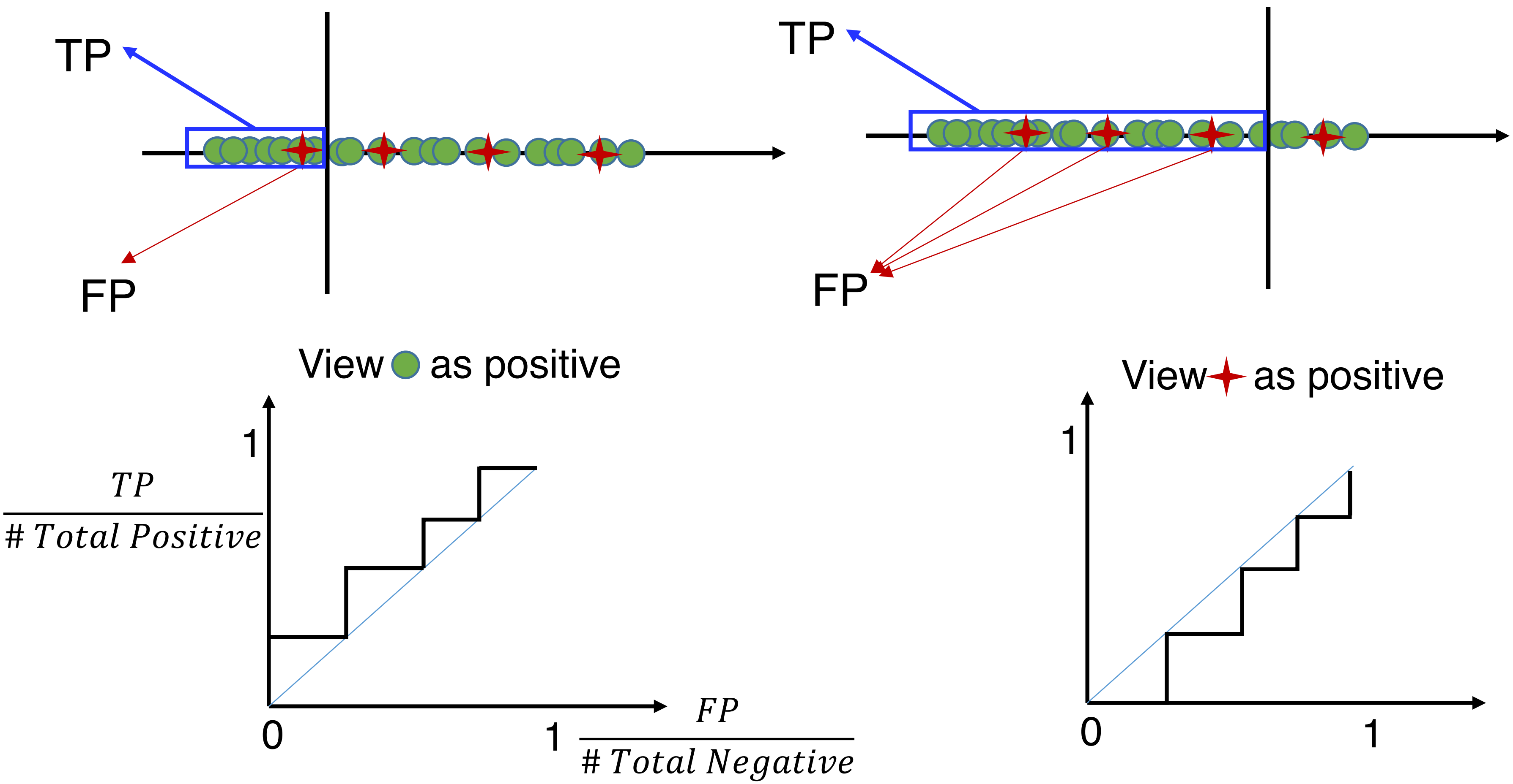
		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

- Accuracy is not sufficient to reflect how well a detector performs
  - For example, if # normal >> # anomalies, detection accuracy is easy to get very high by deeming all testing instances as normal, but it is actually a very bad anomaly detector

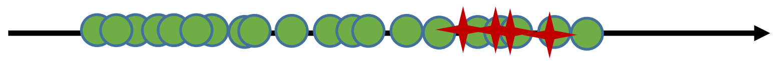


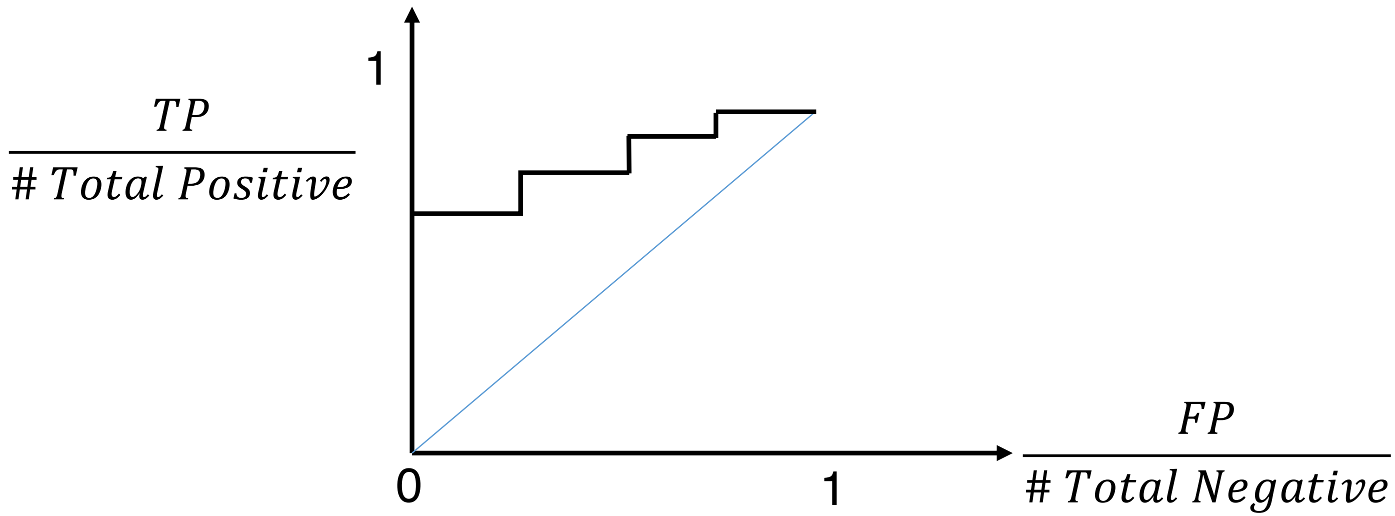
- Receiver operating characteristic (ROC) curve
  - Curve about $\frac{FP}{\#\,Total\ Negative}$ vs $\frac{TP}{\#\,Total\ Positive}$

➢ Example 1 on ROC curve

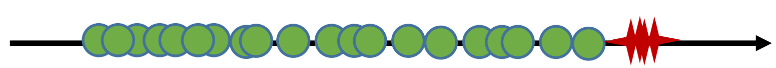


Example 2 on ROC curve



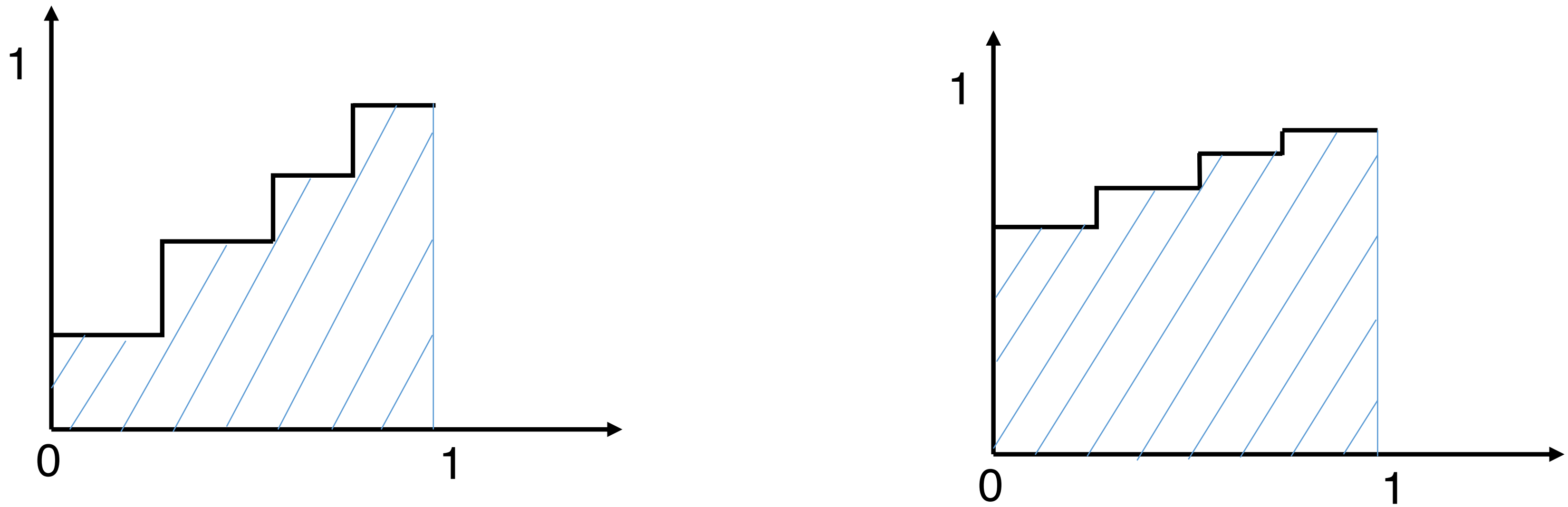


Example 3 on ROC curve



???

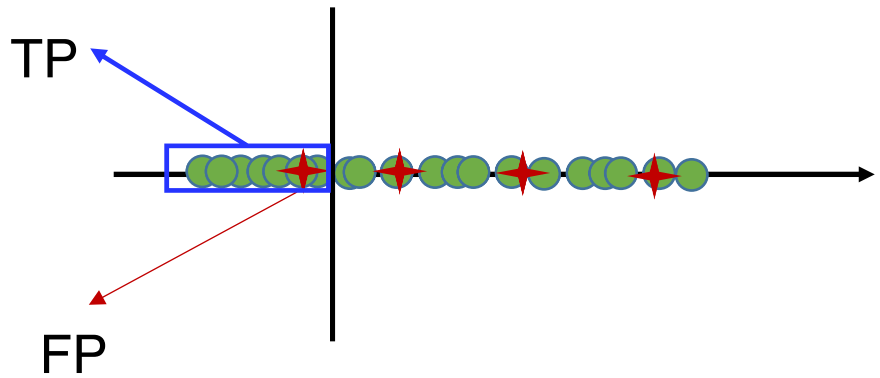
- *Area under ROC (AUROC)* is a good criteria to evaluate how well the normal points are separated from the anomalies



What is the range of the value of AUROC?

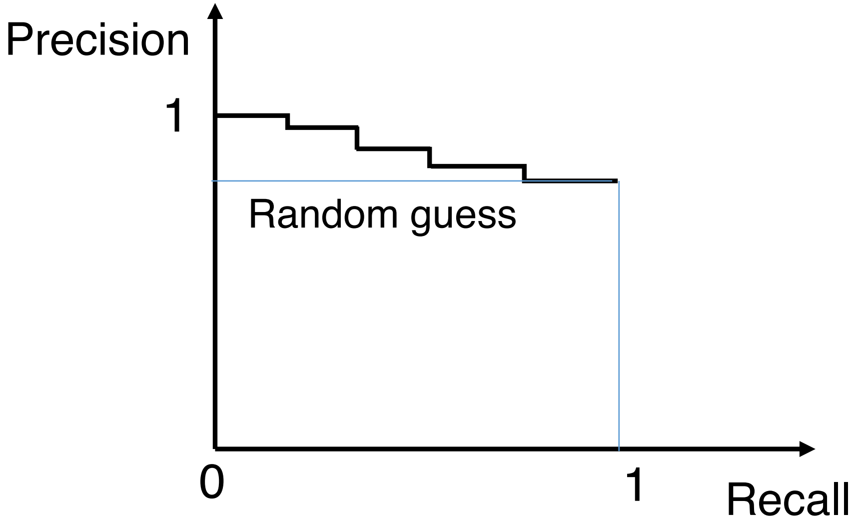
Precision-Recall Curve

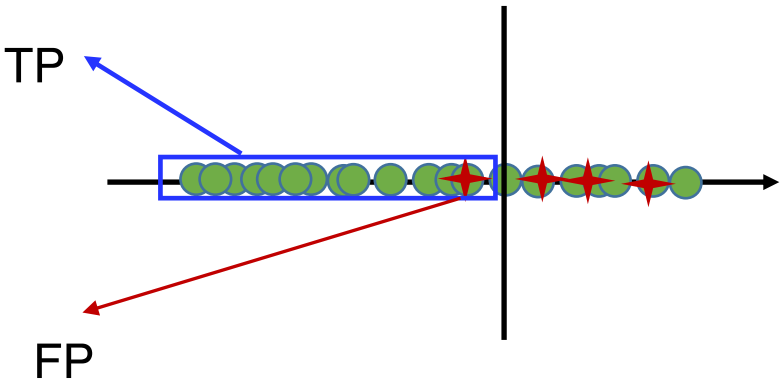
- Another commonly used criteria to measure the performance of anomaly detector is the precision-recall (PR) curve
  - x-axis: $recall = \frac{TP}{\#\,Total\ Positive}$ $\left(= \frac{TP}{all\ green}\right)$



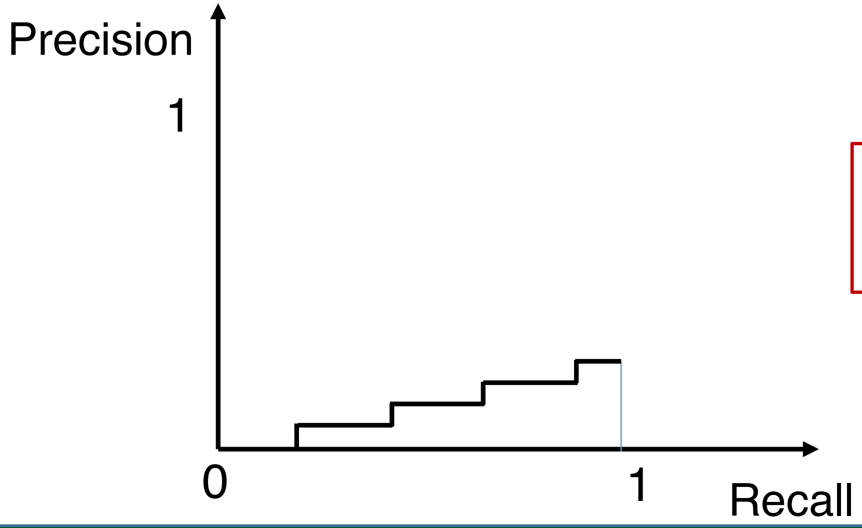
  - y-axis: $precision = \frac{TP}{\#\,Total\ Retrieved}$ $\left(= \frac{TP}{TP + FP}\right)$

- PR curve example (taking *green* as positive)





- PR curve example (taking *red* as positive)



Choosing which points as positive matters!

- Similarly, we can employ the area under PR curve to measure the performance (AUPR)
  - Unlike AUROC, whose value is 0.5 for random guess detector, the AUPR value could be large for random guess detector, *e.g.*, when the positive instances are dominant
  - That is, for random guess detector, its AUROC is always 0.5, but its AUPR could vary in a wide range
  - So, a detector cannot be judged simply according to its absolute value of AUPR. We should compare it with the AUPR of random guess detector
- The F_1 score is also a widely used criteria, which is computed as

$$F_1 = \frac{precision \cdot recall}{0.5 \cdot (precision + recall)}$$

Rough interpretation: if $F_1 * 100$ percent positives is recalled, the precision can reach F_1