Anomaly Detection

Adapted from slides by
Jing Gao
SUNY Buffalo

Definition of Anomalies

- Anomaly is a pattern in the data that does not conform to the expected behavior
- Also referred to as outliers, exceptions, peculiarities, surprise, etc.
- Anomalies translate to significant (often critical) real life entities
  - Cyber intrusions
  - Credit card fraud

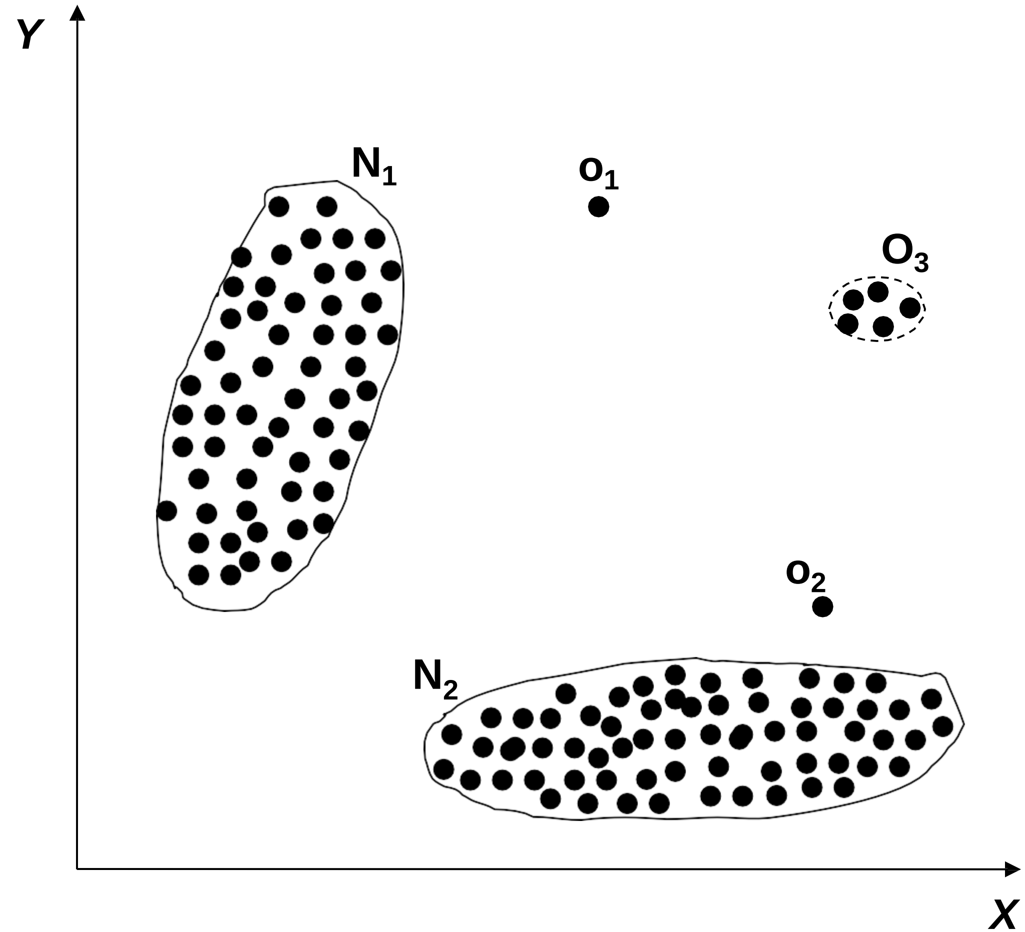
Real World Anomalies

- Credit Card Fraud
  - An abnormally high purchase made on a credit card

- Cyber Intrusions
  - Computer virus spread over Internet

Simple Example

- N_1 and N_2 are regions of normal behavior
- Points o_1 and o_2 are anomalies
- Points in region O_3 are anomalies



Related problems

- Rare Class Mining
- Chance discovery
- Novelty Detection
- Exception Mining
- Noise Removal

Key Challenges

- Defining a representative normal region is challenging
- The boundary between normal and outlying behavior is often not precise
- The exact notion of an outlier is different for different application domains
- Limited availability of labeled data for training/validation
- Malicious adversaries
- Data might contain noise
- Normal behaviour keeps evolving

Aspects of Anomaly Detection Problem

- Nature of input data
- Availability of supervision
- Type of anomaly: point, contextual, structural
- Output of anomaly detection
- Evaluation of anomaly detection techniques

Data Labels

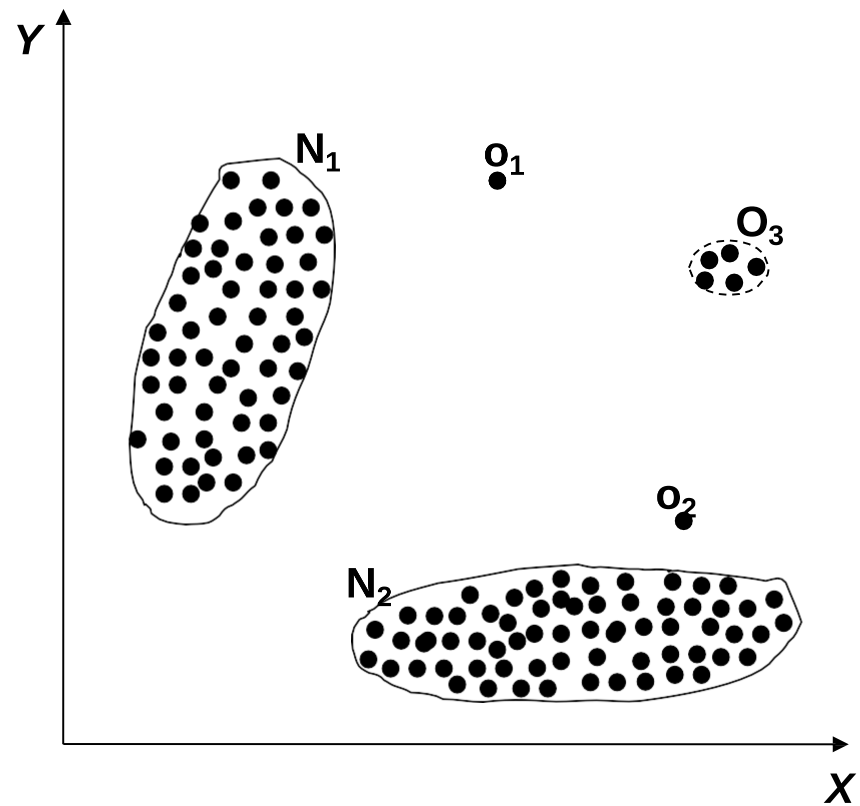
- **Supervised Anomaly Detection**
  - Labels available for both normal data and anomalies
  - Similar to skewed (imbalanced) classification
- **Semi-supervised Anomaly Detection**
  - Limited amount of labeled data
  - Combine supervised and unsupervised techniques
- **Unsupervised Anomaly Detection**
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

Type of Anomalies

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies

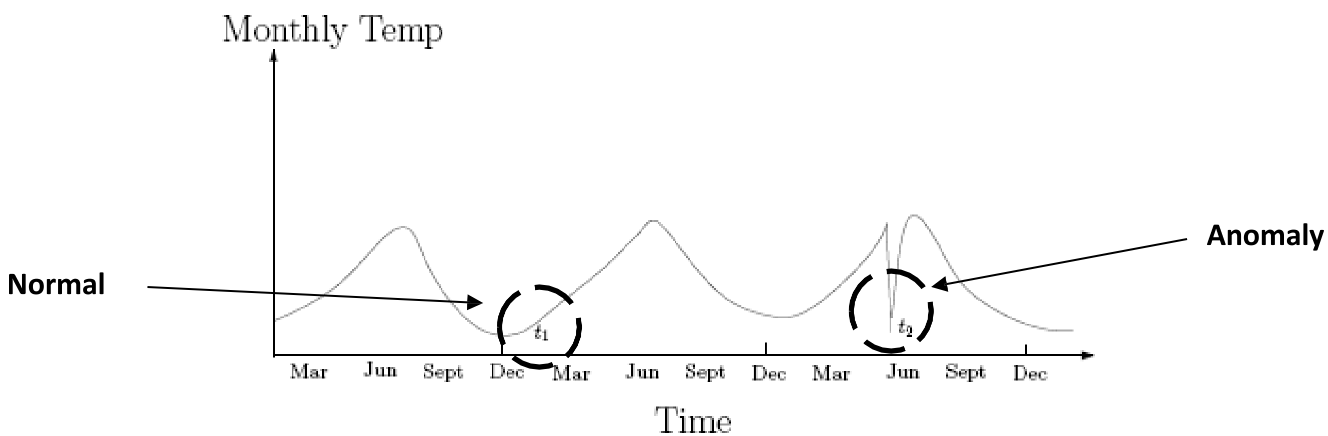
Point Anomalies

- An individual data instance is anomalous w.r.t. the data



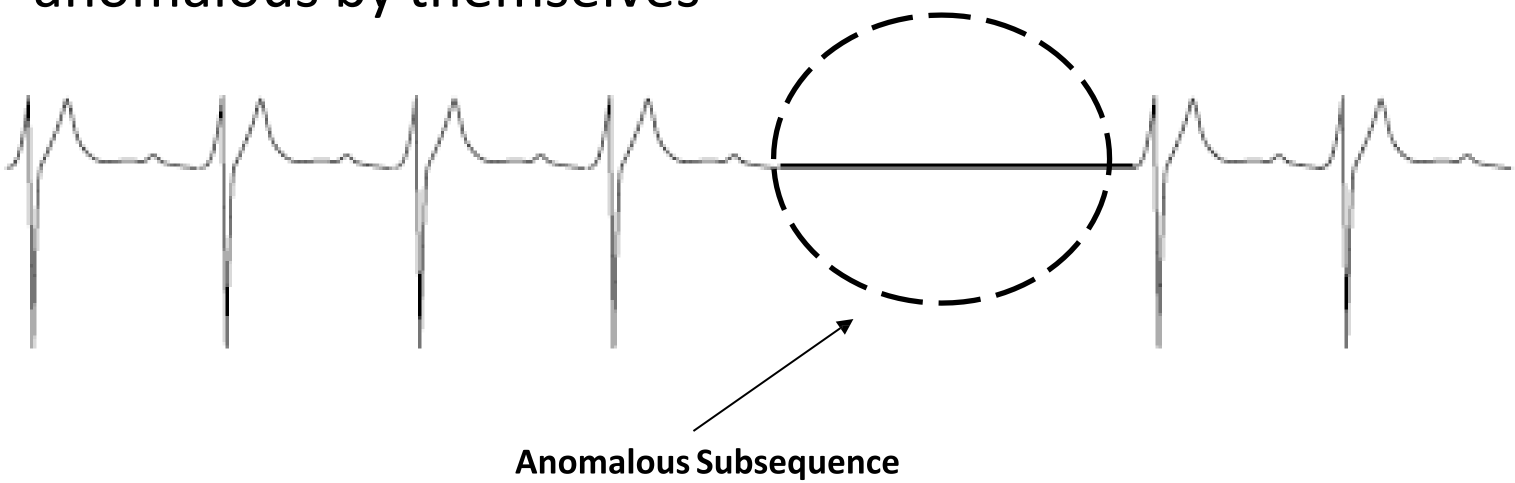
Contextual Anomalies

- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies



Collective Anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential Data
  - Spatial Data
  - Graph Data
- The individual instances within a collective anomaly are not anomalous by themselves



Output of Anomaly Detection

- **Label**
  - Each test instance is given a *normal* or *anomaly* label
  - This is especially true of classification-based approaches
- **Score**
  - Each test instance is assigned an anomaly score
    - Allows the output to be ranked
    - Requires an additional threshold parameter

Metrics for Performance Evaluation

	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	a (TP)	b (FN)
	-	c (FP)	d (TN)

- Measure used in classification:

$$\text{Accuracy} = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$

Limitation of Accuracy

- Anomaly detection
  - Number of negative examples = 9990
  - Number of positive examples = 10

- If model predicts everything to be class 0, accuracy is 9990/10000 = 99.9 %
  - Accuracy is misleading because model does not detect any positive examples

Cost Matrix

	PREDICTED CLASS		
ACTUAL CLASS	C(i|j)	+	-
	+	C(+|+)	C(-|+)
	-	C(+|-)	C(-|-)

C(i|j): Cost of misclassifying class j example as class i

Computing Cost of Classification

Cost Matrix	PREDICTED CLASS		
	C(i|j)	+	-
ACTUAL CLASS	+	-1	100
	-	1	0

Model M_1	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	150	40
	-	60	250

Model M_2	PREDICTED CLASS		
		+	-
ACTUAL CLASS	+	250	45
	-	5	200

Accuracy = 80%

Cost = 3910

Accuracy = 90%

Cost = 4255

Cost-Sensitive Measures

$$\text{Precision (p)} = \frac{a}{a+c}$$

$$\text{Recall (r)} = \frac{a}{a+b}$$

$$\text{F-measure (F)} = \frac{2rp}{r+p} = \frac{2a}{2a+b+c}$$

$$\text{Weighted Accuracy} = \frac{w_1 a + w_4 d}{w_1 a + w_2 b + w_3 c + w_4 d}$$

ROC (Receiver Operating Characteristic)

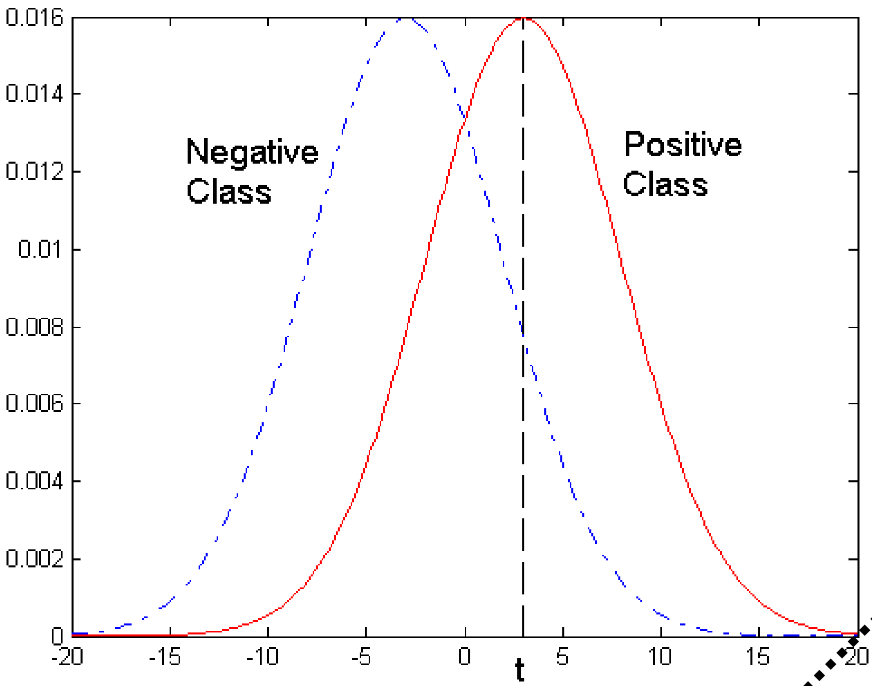
- ROC curve plots TPR (Recall) on the y-axis against FPR (FP/#N) on the x-axis

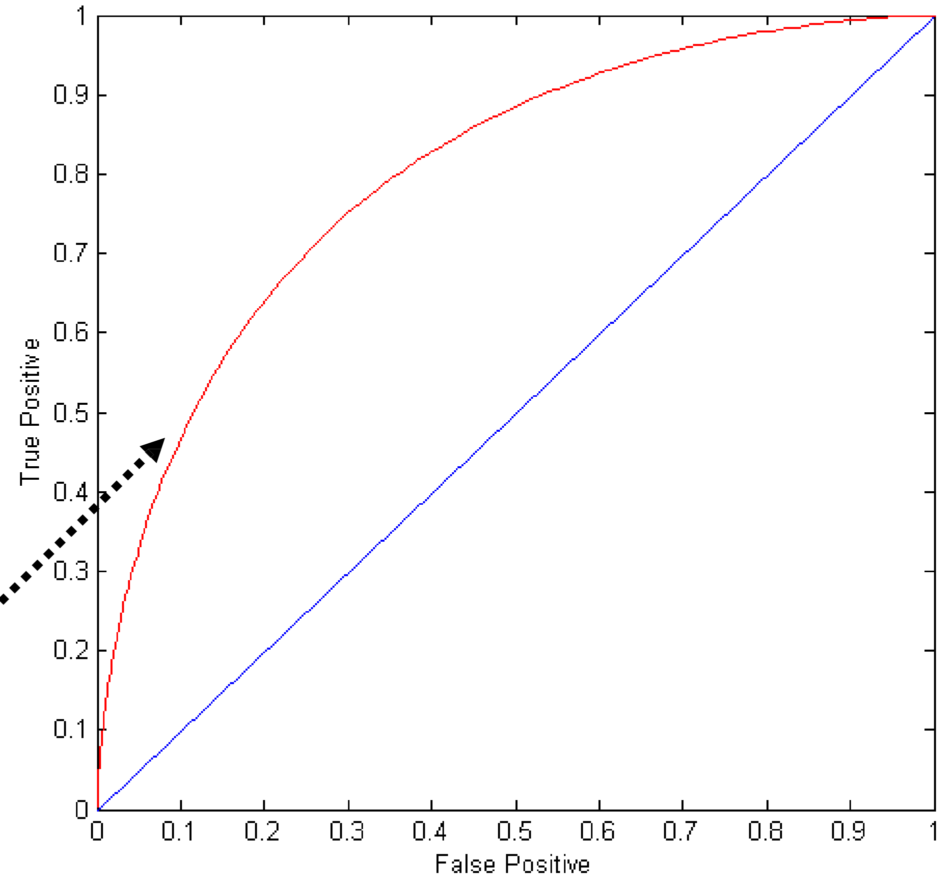
- Performance of each classifier represented as a point on the ROC curve
  - changing the threshold of algorithm, sample distribution or cost matrix changes the location of the point

ROC Curve

- 1-dimensional data set containing 2 classes (positive and negative)

- any points located at x > t is classified as positive





At threshold t:

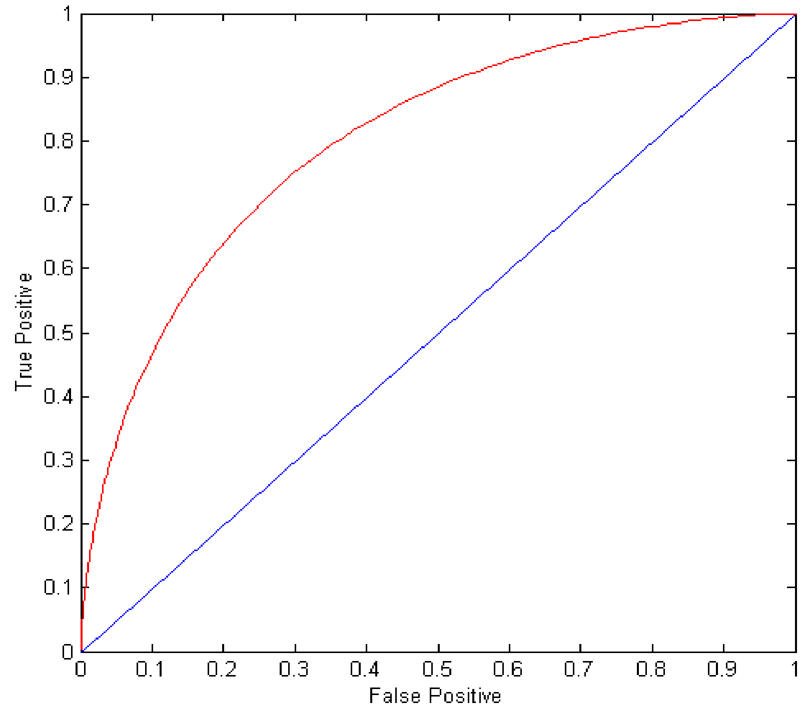
TP=0.5, FN=0.5, FP=0.12, FN=0.88

ROC Curve

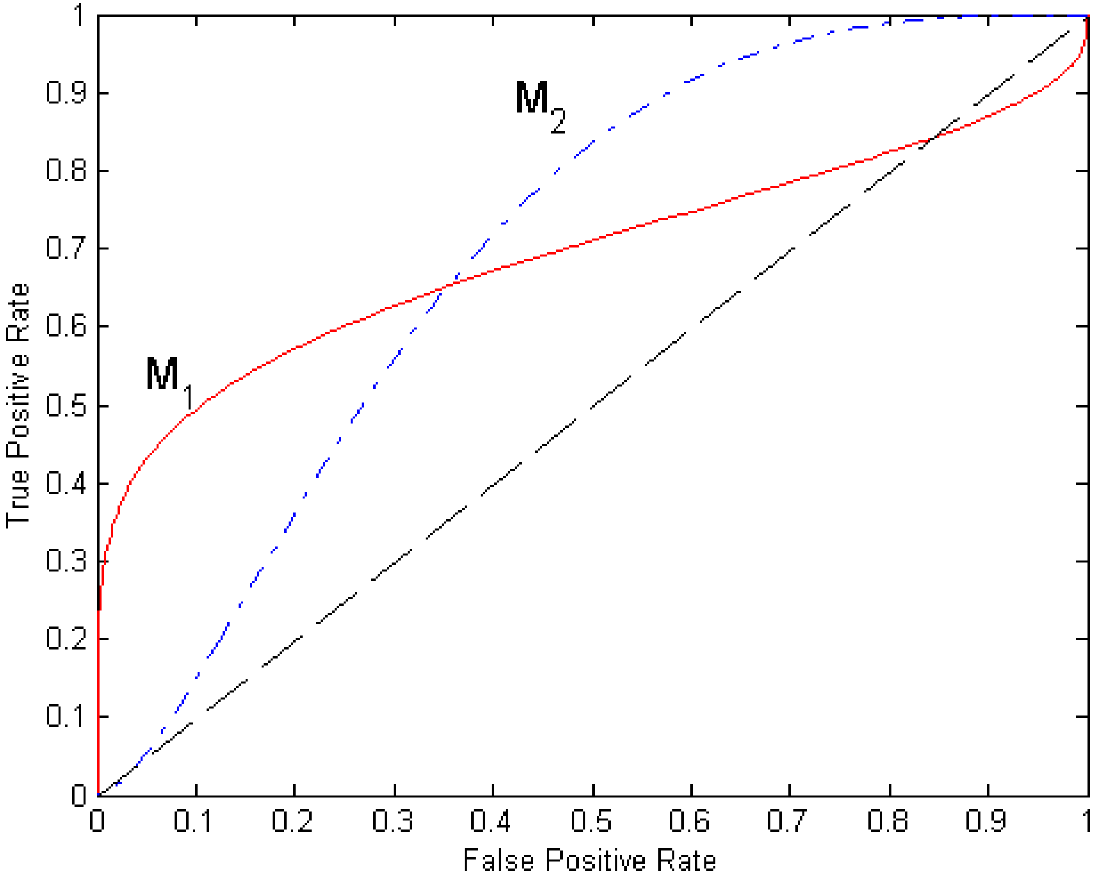
(TPR,FPR):

- (0,0): declare everything
  to be negative class
- (1,1): declare everything
  to be positive class
- (1,0): ideal

- Diagonal line:
  - Random guessing
  - Below diagonal line:
    - prediction is opposite of the true class



Using ROC for Model Comparison



- Comparing two models
  - M_1 is better for small FPR
  - M_2 is better for large FPR
- Area Under the ROC curve
  - Ideal:
    - Area = 1
  - Random guess:
    - Area = 0.5

How to Construct an ROC curve

Instance	Score	Label
1	0.95	+
2	0.93	+
3	0.87	-
4	0.85	-
5	0.85	-
6	0.85	+
7	0.76	-
8	0.53	+

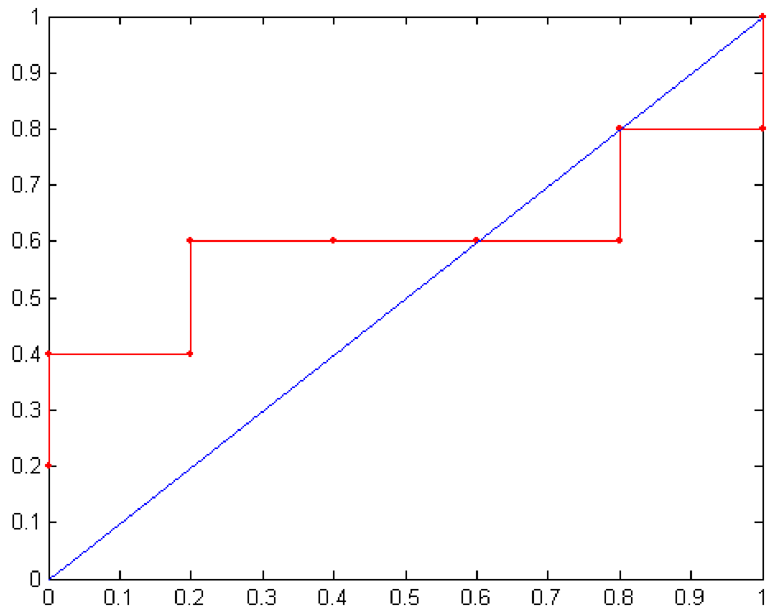
	PREDICTED CLASS		
ACTUAL CLASS		+	-
	+	a (TP)	b (FN)
	-	c (FP)	d (TN)

- Calculate the outlier scores of the given instances
- Sort the instances according to the scores in decreasing order
- Apply threshold at each unique value of the score
- Count the number of TP, FP, TN, FN at each threshold
- TP rate, TPR = TP/(TP+FN)
- FP rate, FPR = FP/(FP + TN)

How to construct an ROC curve

Class	+	-	+	-	-	-	+	-	+	+	
Threshold >=	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

ROC Curve:



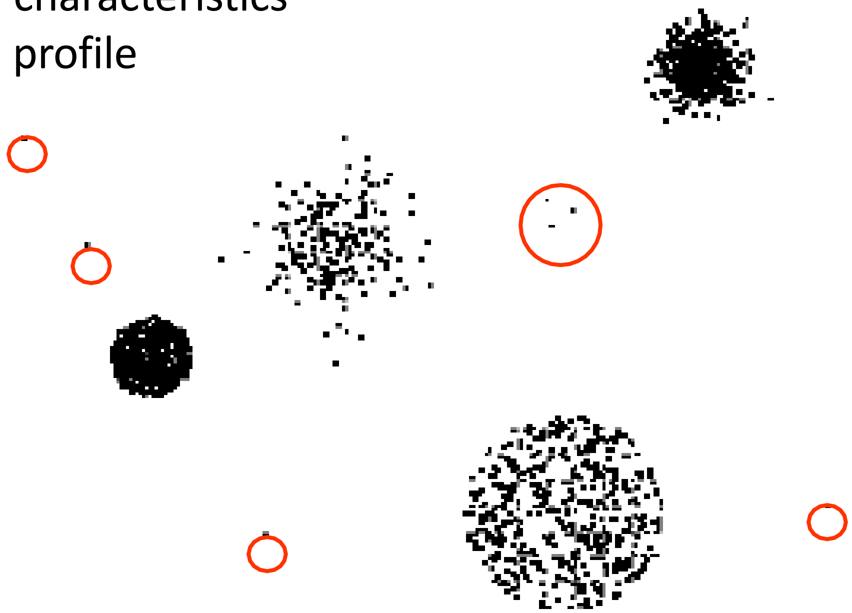
Area under ROC curve = prob. that a randomly sample positive example will score higher than a randomly sampled negative example

Applications of Anomaly Detection

- Network intrusion detection
- Insurance / Credit card fraud detection
- Healthcare Informatics / Medical diagnostics
- Image Processing / Video surveillance
- …

Anomaly Detection Schemes

- **General Steps**
  - Build a profile of the “normal” behavior
    - Profile can be patterns or summary statistics for the overall population
  - Use the “normal” profile to detect anomalies
    - Anomalies are observations whose characteristics differ significantly from the normal profile

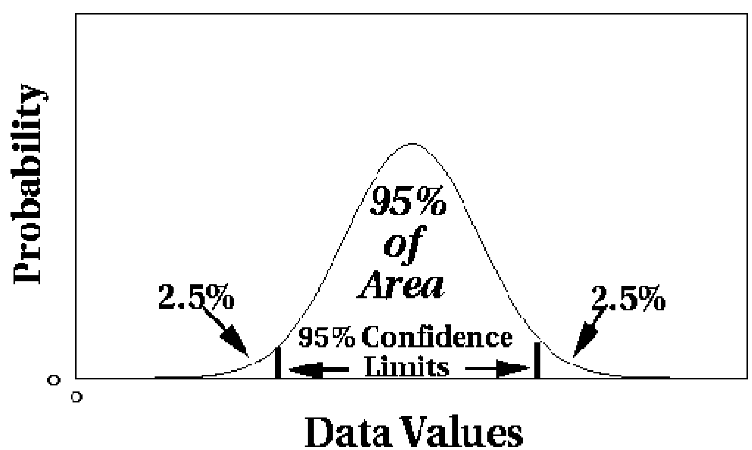


- **Methods**
  - Statistical-based
  - Distance-based
  - Model-based

Statistical Approaches

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)

- Apply a statistical test that depends on
  - Data distribution
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit)



Limitations of Statistical Approaches

- Most of the tests are for a single attribute
- In many cases, data distribution may not be known
- For high dimensional data, it may be difficult to estimate the true distribution

Distance-based Approaches

- Data is represented as a vector of features

- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

Nearest-Neighbor Based Approach

- **Approach:**
  - Compute the distance between every pair of data points

  - There are various ways to define outliers:
    - Data points for which there are fewer than *p* neighboring points within a distance *D*

    - The top *n* data points whose distance to the *k*-th nearest neighbor is greatest

    - The top *n* data points whose average distance to the *k* nearest neighbors is greatest

Distance-Based Outlier Detection

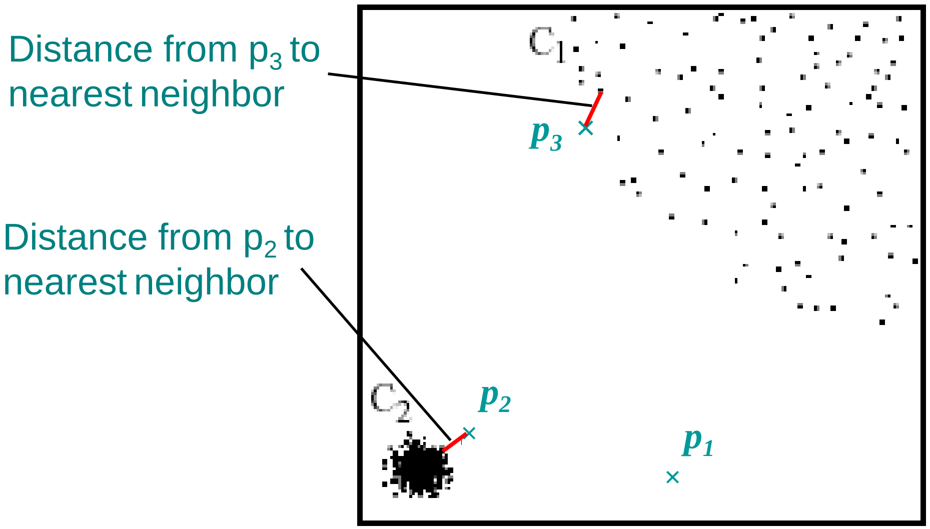
- For each object o, examine the # of other objects in the *r*-neighborhood of o, where *r* is a user-specified **distance threshold**
- An object o is an outlier if most (taking π as a **fraction threshold**) of the objects in D are far away from o, i.e., not in the r-neighborhood of o

- An object o is a DB(r, π) outlier if $$\frac{\|\{o'|dist(o,o') \leq r\}\|}{\|D\|} \leq \pi$$
- Equivalently, one can check the distance between *o* and its *k*-th nearest neighbor o_k, where $k = \lceil \pi \|D\| \rceil$. *o* is an outlier if dist(o, o_k) > r

Density-based Approach

- For each point, compute the density of its local neighborhood
- Points whose local density is significantly lower than its nearest neighbor‘s local density are consider outliers

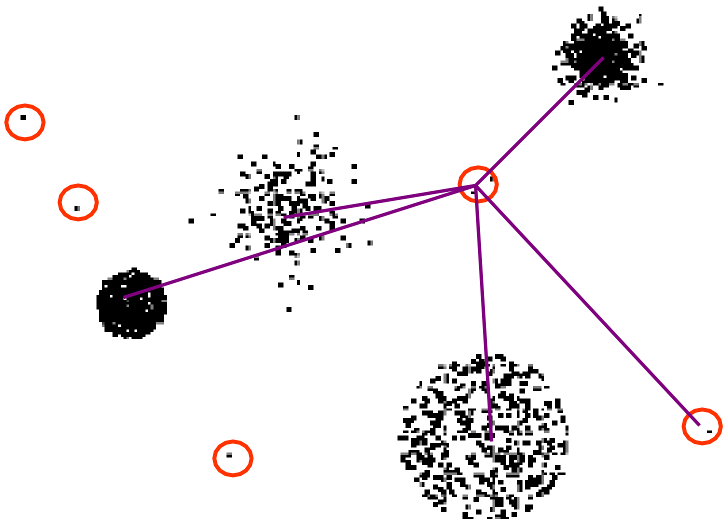
Example:



In the *NN* approach, p_2 is not considered as outlier, while a density based approach may find both p_1 and p_2 as outliers

Clustering-Based

- **Basic idea:**
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, they are outliers

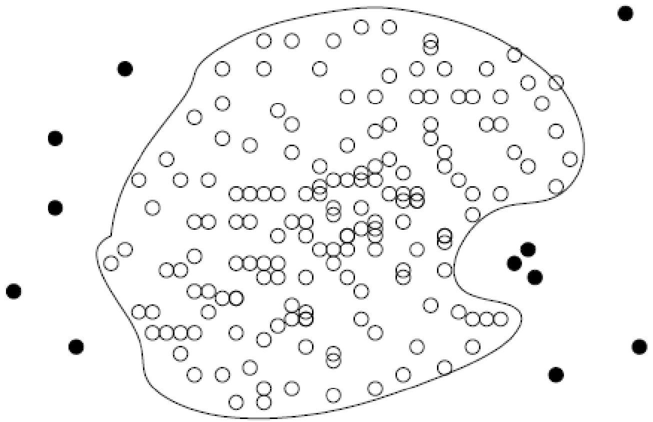


Classification-Based Methods

- Idea: Train a classification model that can distinguish “normal” data from outliers
- Consider a training set that contains samples labeled as “normal” and others labeled as “outlier”
  - But, the training set is typically heavily biased: # of “normal” samples likely far exceeds # of outlier samples
- Handle the imbalanced distribution
  - Oversampling positives and/or undersampling negatives
  - Alter decision threshold
  - Cost-sensitive learning

One-Class Model

- One-class model: A classifier is built to describe only the normal class
  - Learn the decision boundary of the normal class using classification methods such as SVM
  - Any samples that do not belong to the normal class (not within the decision boundary) are declared as outliers
  - Adv: can detect new outliers that may not appear close to any outlier objects in the training set



Take-away Message

- Definition of outlier detection
- Applications of outlier detection
- Evaluation of outlier detection techniques
- Unsupervised approaches (statistical, distance, density-based)