CS 584
Data Mining

Anomaly Detection

What are Anomalies?

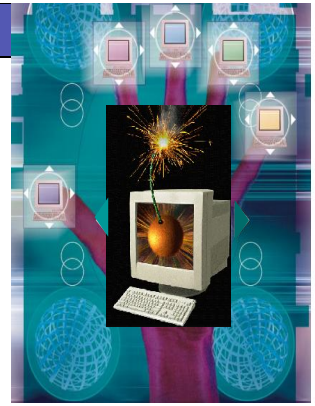
- Anomaly is a pattern in the data that does not conform to the expected behavior
- Also referred to as outliers, exceptions, peculiarities, surprise, etc.
- Anomalies translate to significant (often critical) real life entities
  - Cyber intrusions
  - Credit card fraud

Real World Anomalies

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

- Cyber Intrusions
  - A web server involved in *ftp* traffic

Intrusion Detection



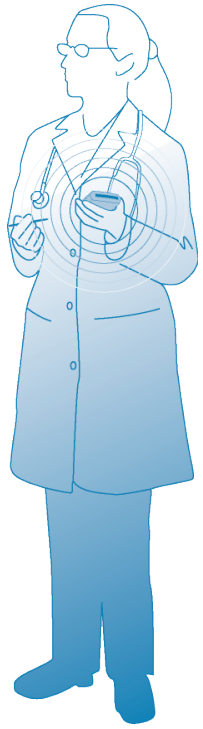
- Intrusion Detection:
  - Process of monitoring the events occurring in a computer system or network and analyzing them for intrusions
  - Intrusions are defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional signature-based intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

Fraud Detection

- Fraud detection refers to detection of criminal activities occurring in commercial organizations
  - Malicious users might be the actual customers of the organization or might be posing as a customer (also known as identity theft).
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading
- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high

Healthcare Informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key Challenges
  - Only normal labels available
  - Misclassification cost is very high



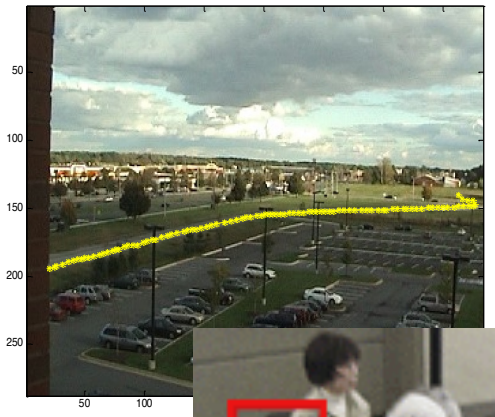
Industrial Damage Detection

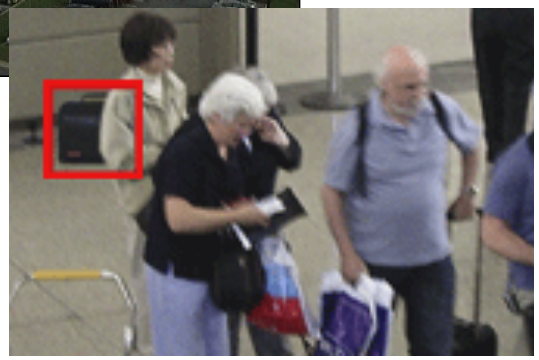
- Industrial damage detection refers to detection of different faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
  - Example: Aircraft Safety
    - Anomalous Aircraft (Engine) / Fleet Usage
    - Anomalies in engine combustion data
    - Total aircraft health and usage management

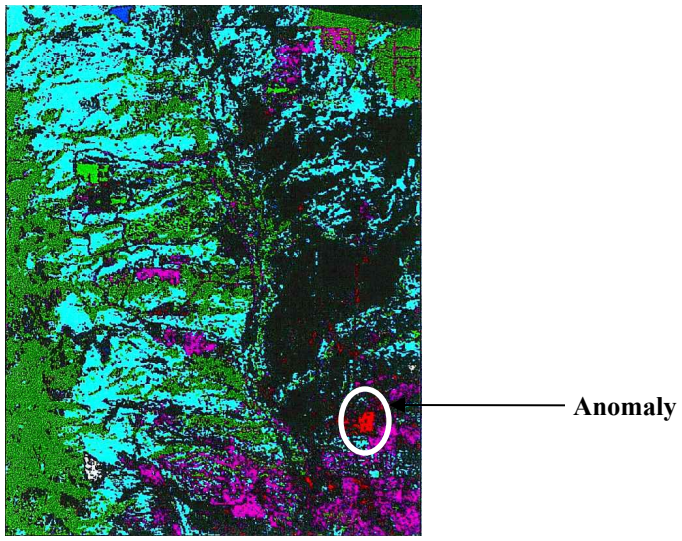
- Key Challenges
  - Data is extremely huge, noisy and unlabelled
  - Most of applications exhibit temporal behavior
  - Detecting anomalous events typically require immediate intervention

Image Processing

- Detecting outliers in an image monitored over time
- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Data sets are very large

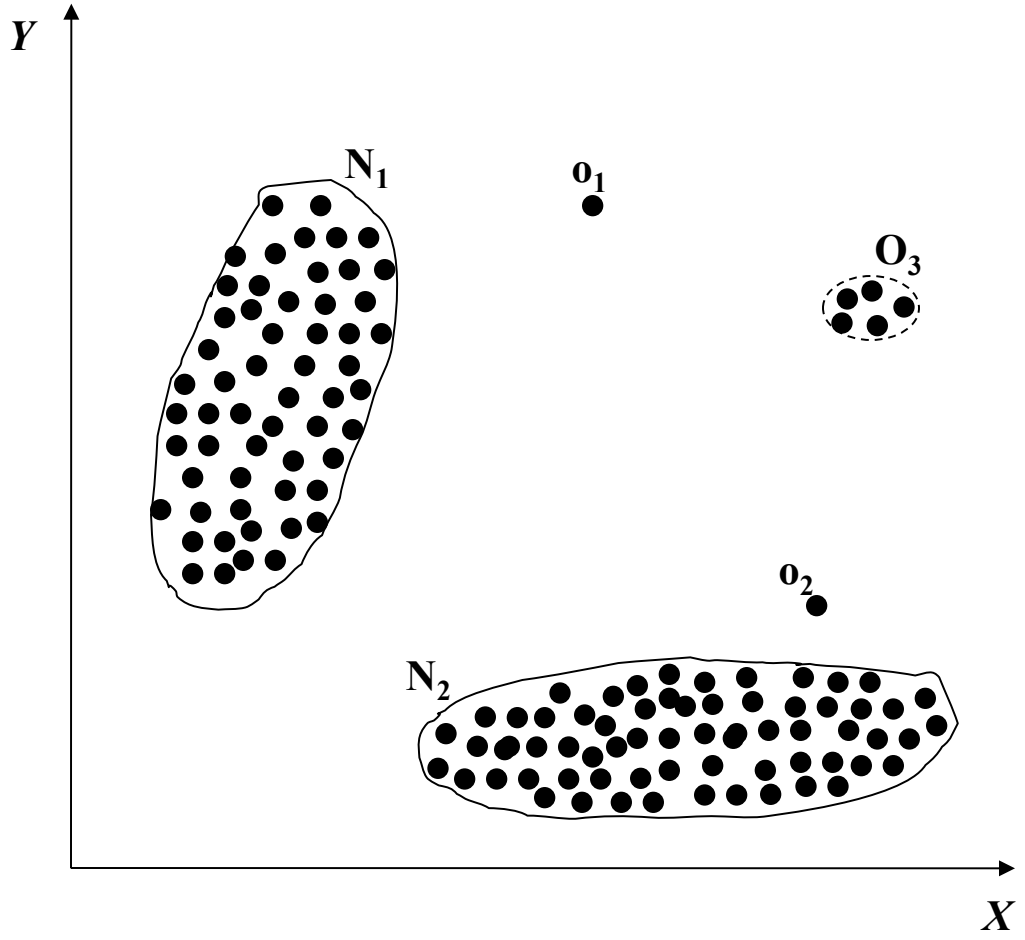






Simple Example

- 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



Key Challenges

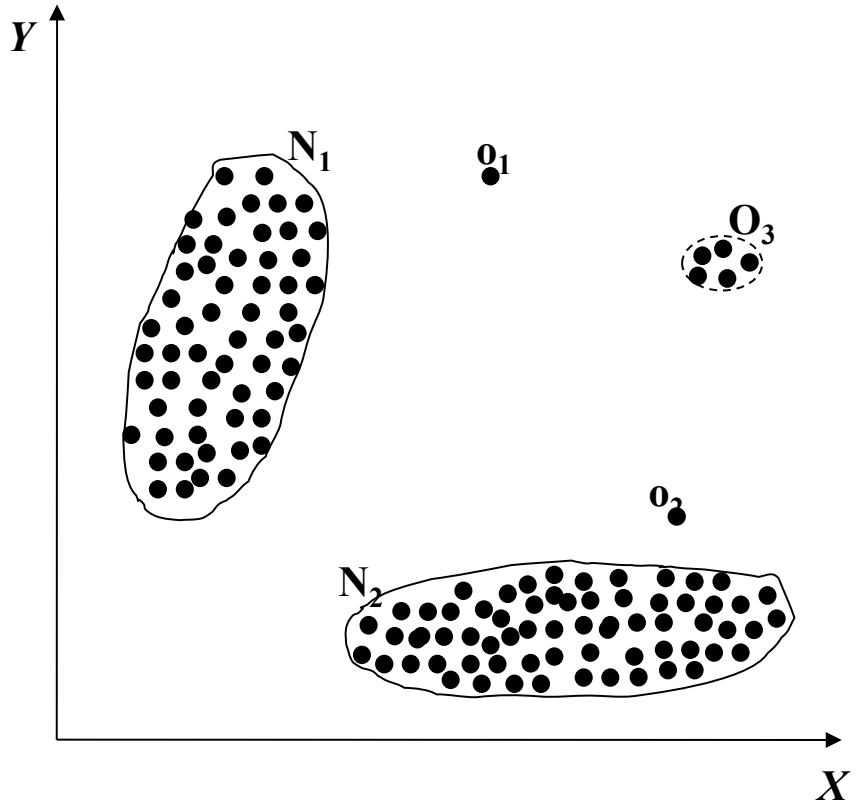
- Defining a representative normal region is challenging
- The boundary between normal and outlying behavior is often not precise
- The exact notion of an outlier is different for different application domains
- Availability of labeled data for training/validation
- Malicious adversaries
- Data might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

Type of Anomaly

- Point Anomalies
- Contextual Anomalies
- Collective Anomalies

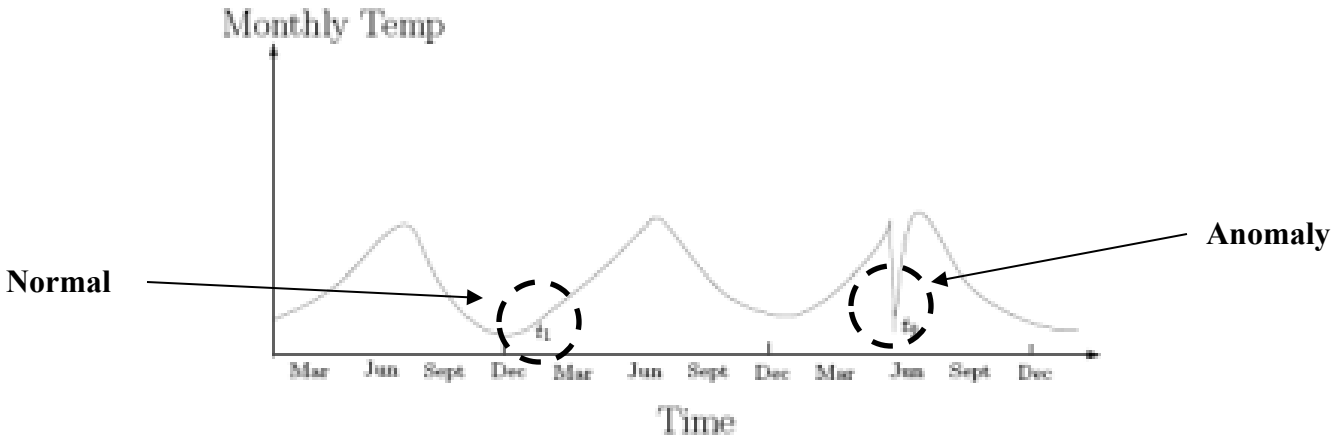
Point Anomalies

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



Contextual Anomalies

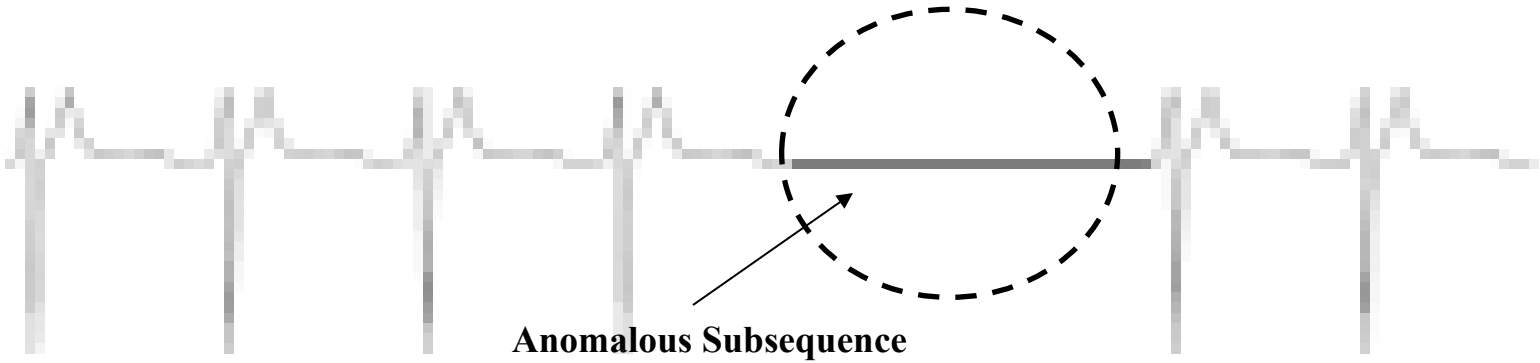
- An individual data instance is anomalous within a context
- Requires a notion of context
- Also referred to as conditional anomalies*



*** Xiuyao Song, Mingxi Wu, Christopher Jermaine, Sanjay Ranka, Conditional Anomaly Detection, IEEE Transactions on Data and Knowledge Engineering, 2006.**

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



Class Label Availability

How would you do this (3 groups)

- Supervised
- Semi-supervised
- Unsupervised

Think about a method, as well as evaluation plan.

Evaluation of Anomaly Detection – F-value

- Accuracy is not sufficient metric for evaluation
  - Example: network traffic data set with 99.9% of normal data and 0.1% of intrusions
  - Trivial classifier that labels everything with the normal class can achieve 99.9% accuracy !!!!!

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

anomaly class – C

normal class – NC

- **Focus on both recall and precision**
  - **Recall (R)= TP/(TP + FN)**
  - **Precision (P) = TP/(TP + FP)**
- **F – measure = 2*R*P/(R+P)**

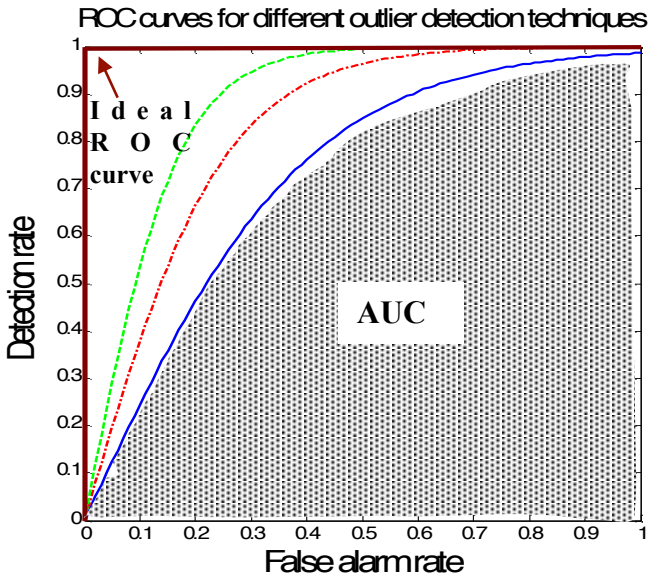
Evaluation of Outlier Detection – ROC & AUC

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

anomaly class – C

normal class – NC

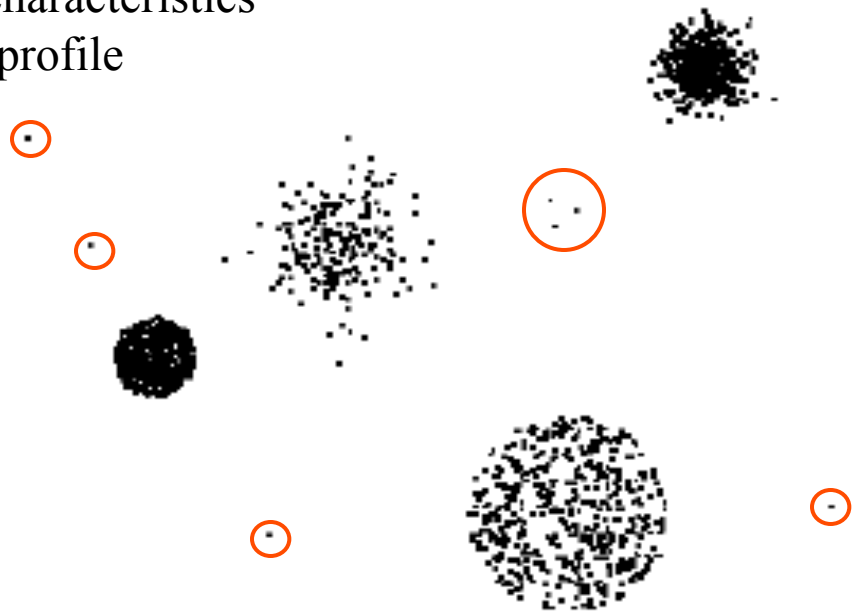
- Standard measures for evaluating anomaly detection problems:
  - *Recall (Detection rate)* - ratio between the number of correctly detected anomalies and the total number of anomalies
  - *False alarm (false positive) rate* – ratio between the number of data records from normal class that are misclassified as anomalies and the total number of data records from normal class
  - *ROC Curve* is a trade-off between detection rate and false alarm rate
  - *Area under the ROC curve (AUC)* is computed using a trapezoid rule



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

- Types of anomaly detection schemes
  - Graphical & Statistical-based
  - Distance-based/Density-based
  - Model-based



Classification Based Techniques

- Main idea: build a classification model for normal (and anomalous (rare)) events based on labeled training data, and use it to classify each new unseen event
- Classification models must be able to handle skewed (imbalanced) class distributions
- Categories:
  - *Supervised classification techniques*
    - Require knowledge of both ***normal*** and ***anomaly*** class
    - Build classifier to distinguish between normal and known anomalies
  - *Semi-supervised classification techniques*
    - Require knowledge of ***normal*** class only!
    - Use modified classification model to learn the normal behavior and then detect any deviations from normal behavior as anomalous

Classification Based Techniques

- Advantages:
  - *Supervised classification techniques*
    - Models that can be easily understood
    - High accuracy in detecting many kinds of known anomalies
  - *Semi-supervised classification techniques*
    - Models that can be easily understood
    - Normal behavior can be accurately learned
- Drawbacks:
  - *Supervised classification techniques*
    - Require both labels from both normal and anomaly class
    - Cannot detect unknown and emerging anomalies
  - *Semi-supervised classification techniques*
    - Require labels from normal class
    - Possible high false alarm rate - previously unseen (yet legitimate) data records may be recognized as anomalies

Supervised Classification Techniques

- Manipulating data records (oversampling / undersampling / generating artificial examples)
- Rule based techniques
- Model based techniques
  - Neural network based approaches
  - Support Vector machines (SVM) based approaches
  - Bayesian networks based approaches
- Cost-sensitive classification techniques
- Ensemble based algorithms

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

K-NN based outlier

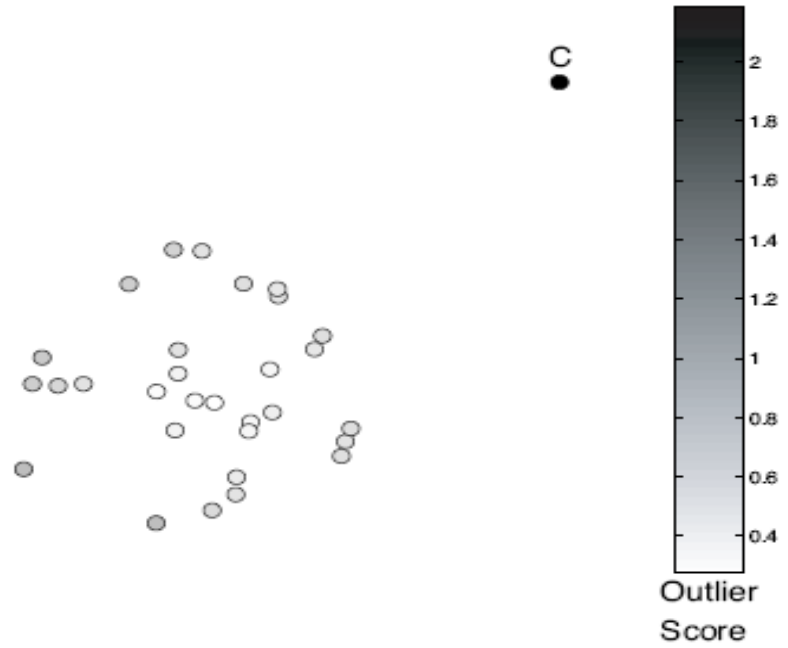


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

Sensitive to “k”, neighbors of outliers

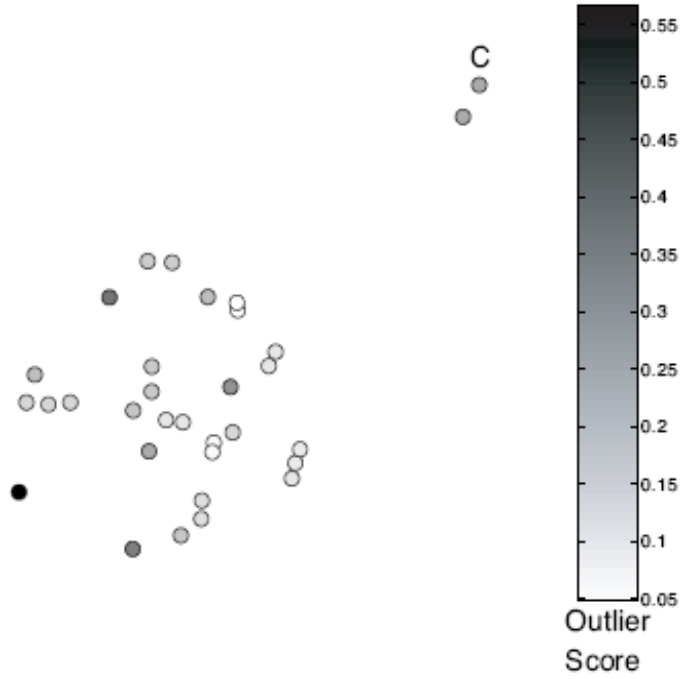


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

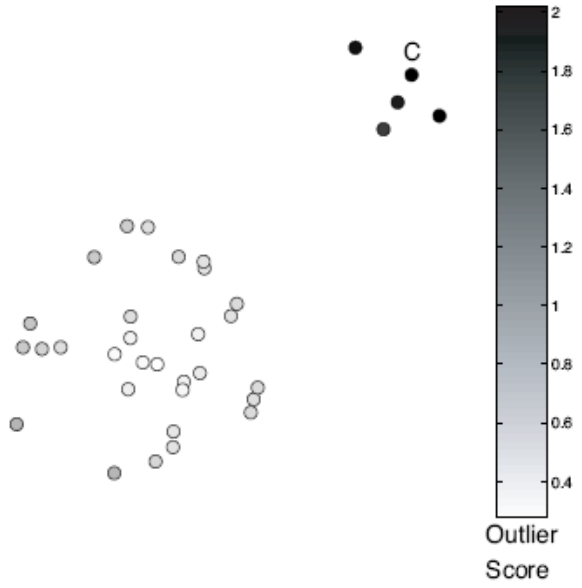


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

Clusters of differing densities

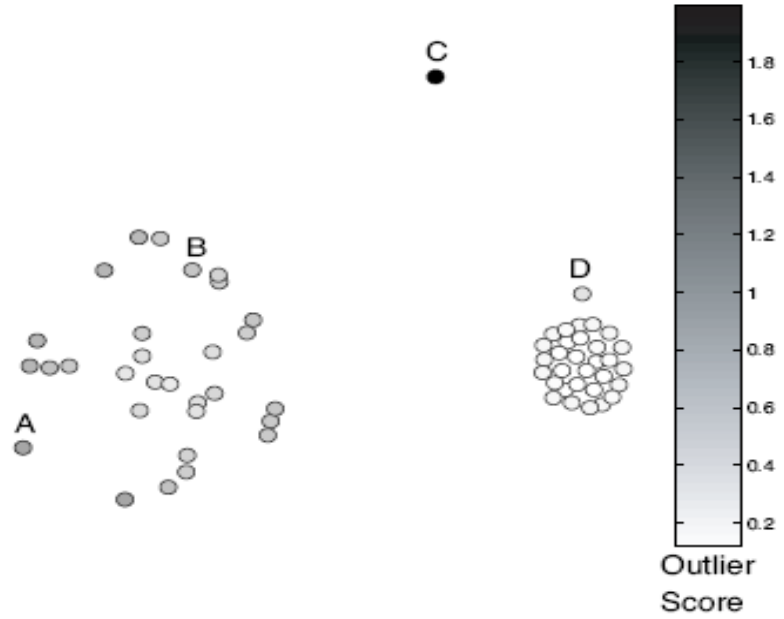
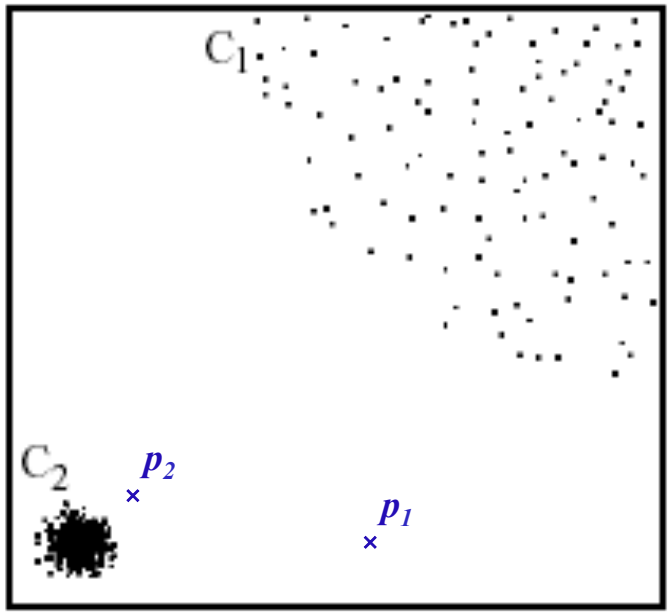


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value



In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

Using density-based measure

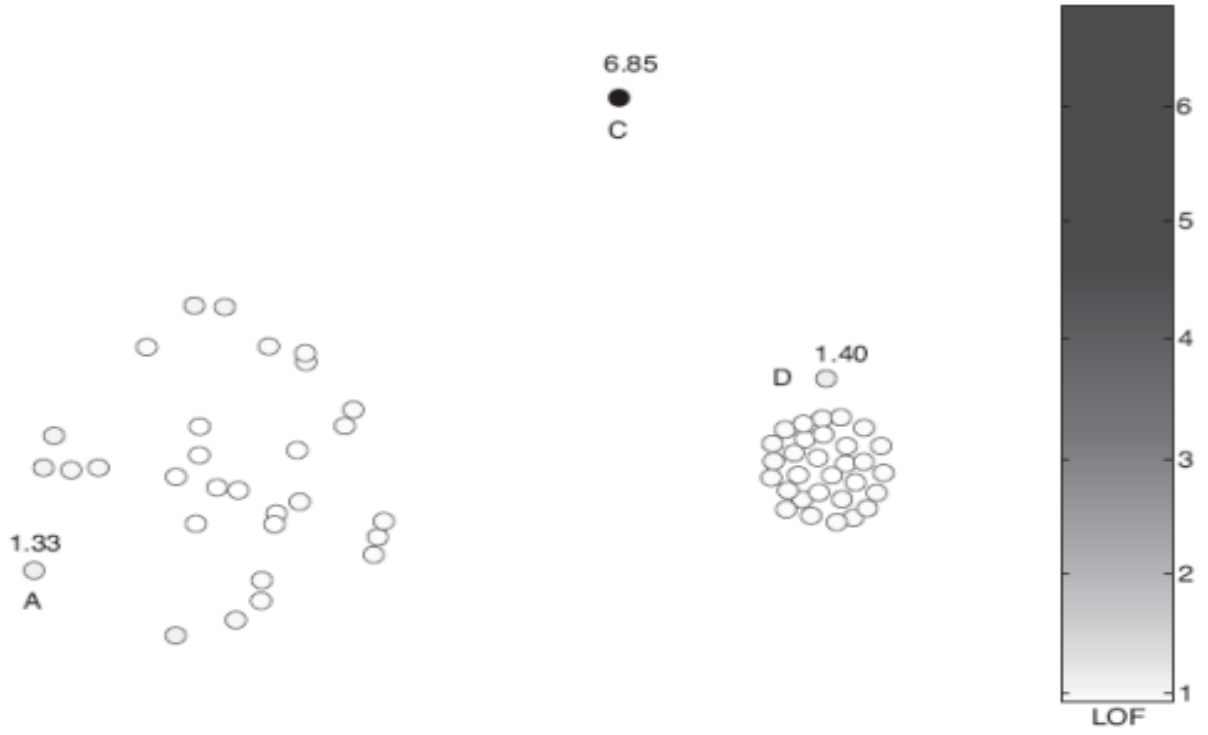


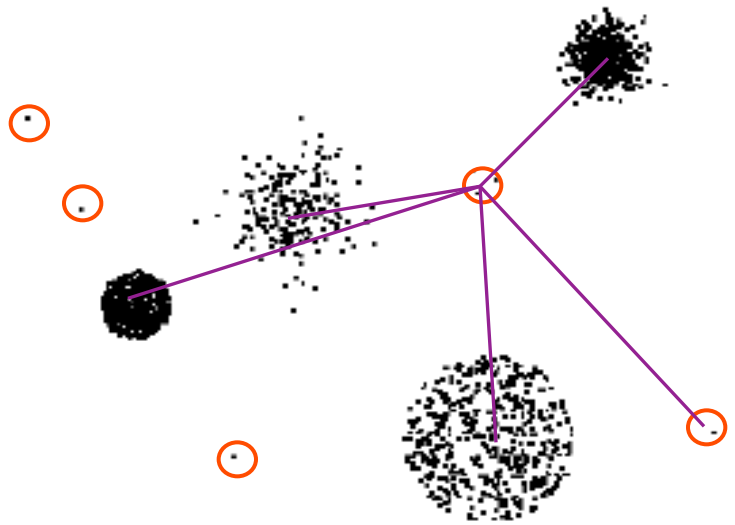
Figure 10.8. Relative density (LOF) outlier scores for two-dimensional points of Figure 10.7.

Strengths and weaknesses

- Density-base provide outlier score
- Complexity like distance-based
  - O(m*m)
- Need to still pick thresholds for upper and lower bounds (e.g. for possible values of k)

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



Distance to Centroid

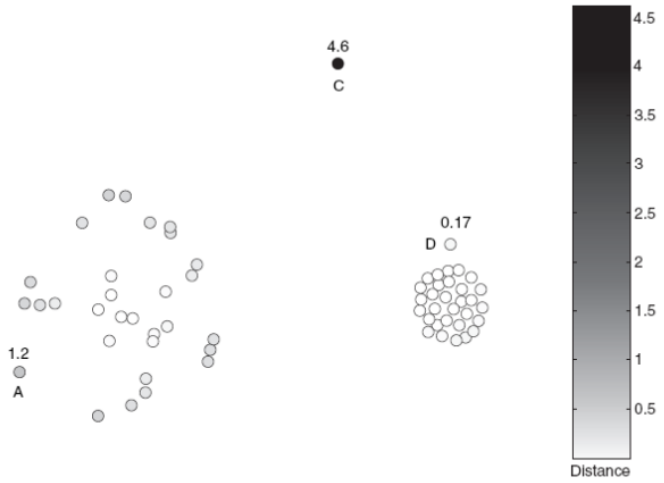


Figure 10.9. Distance of points from closest centroid.

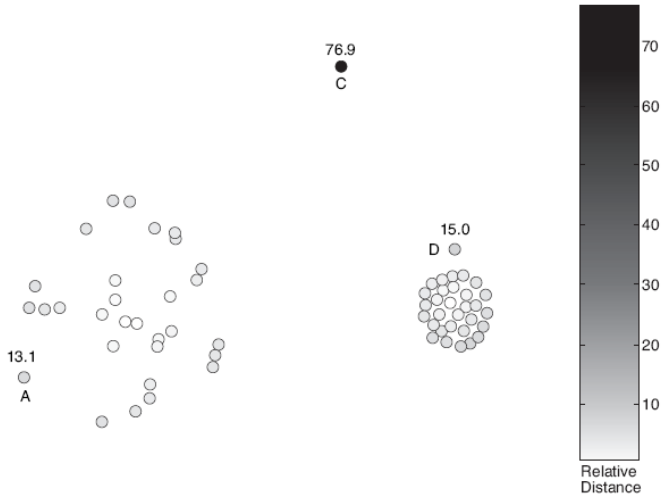


Figure 10.10. Relative distance of points from closest centroid.

Kmeans-based Clustering

- Value of “k”?
- Should we remove outliers and redo the clustering ?