PURDUE UNIVERSITY | Department of Computer Science

CS57300: Data Mining

Anomaly Detection

Prof. Chris Clifton

14 April 2022

Some materials from *Introduction to Data Mining* by Tan, Steinbach and Kumar

PURDUE UNIVERSITY
Department of Computer Science

Task

- **Anomalies/outliers**: data points that are considerably "different" from the remainder of the data
- Variants:
  - Find all points with anomaly scores > threshold
  - Find point with largest anomaly score
  - Given a database D with mostly normal points, compute the anomaly score of a point x with respect to D

PURDUE UNIVERSITY
Department of Computer Science

Examples

- Fraud detection
- Intrusion detection
- Ecosystem disturbances
- System monitoring
- Biosurveillance/public health
- Data preprocessing

PURDUE UNIVERSITY
Department of Computer Science

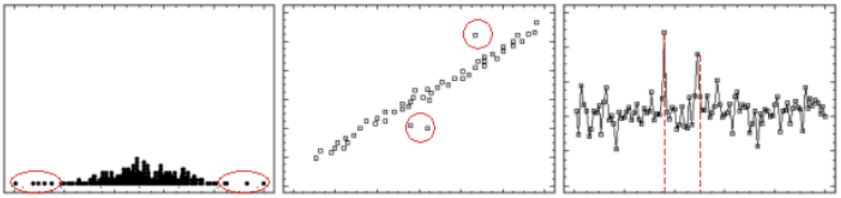
Types of anomalies

- Data from different classes
  - "An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism"
- Natural variation
  - Extreme or unlikely variations are often interesting
- Data measurement and collection errors
  - Preprocess to remove

PURDUE UNIVERSITY
Department of Computer Science

Defining an outlier

- Notion of outlier is highly subjective and domain dependent
- However, most definitions can be viewed as defining a distribution for “normal” data and then looking for deviations from that distribution

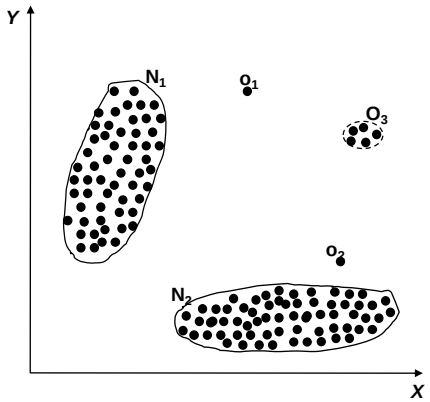


Source: Osmar Zaiane, UAlberta, PKDD

PURDUE UNIVERSITY
Department of Computer Science

Point anomalies

- An **individual** data instance is anomalous with respect to the **data**

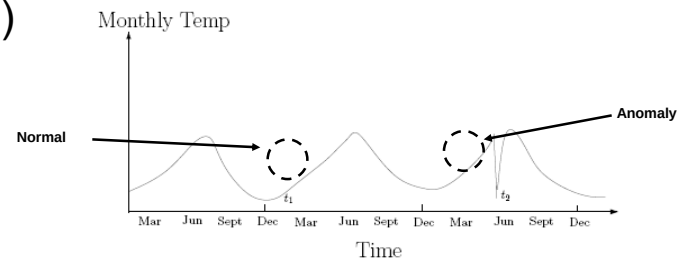


Source: Lazarevic et al, ECML/PKDD’08 Tutorial

PURDUE UNIVERSITY
Department of Computer Science

Contextual anomalies

- An **individual** data instance is anomalous within a **context**
- Requires a notion of context
- Also referred to as conditional anomalies (Song et. al, TDKE ’06)

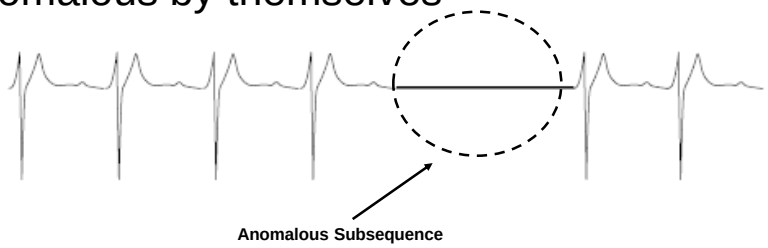


Source: Lazarevic et al, ECML/PKDD’08 Tutorial

PURDUE UNIVERSITY
Department of Computer Science

Collective anomalies

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



Source: Lazarevic et al, ECML/PKDD’08 Tutorial

PURDUE UNIVERSITY
Department of Computer Science

Anomaly detection

- Challenges
  - How many attributes are used to define an outlier?
  - How many outliers are there in the data?
  - Class labels are costly (evaluation can be challenging)
  - Skewed class distribution (finding needles in haystack)
- Working assumption:
  - There are considerably more “normal” observations than “abnormal” observations in the data

PURDUE UNIVERSITY
Department of Computer Science

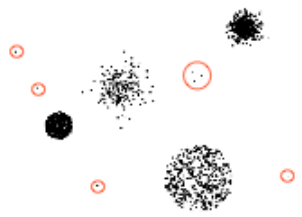
Approaches

- Supervised
  - Labels available for both normal data and anomalies
  - Similar to classification with imbalanced classes
- Semi-supervised
  - Labels available only for normal data
- Unsupervised
  - No labels assumed
  - Based on the assumption that anomalies are very rare compared to normal data

PURDUE UNIVERSITY
Department of Computer Science

Unsupervised (point) anomaly detection

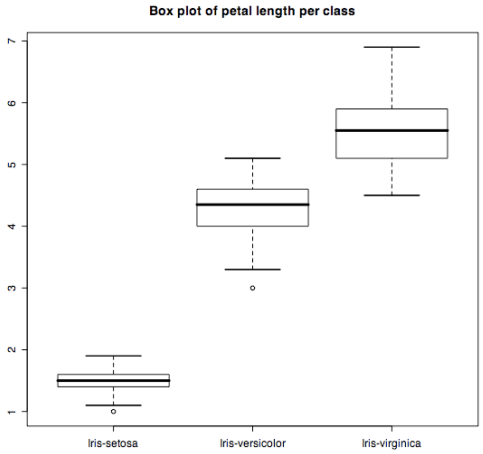
- General method
  - Build a profile of “normal” behavior based on patterns or summary statistics for the overall population
  - Use deviations from “normal” to detect anomalies
- Types of methods
  - Visual and statistical-based
  - Distance-based
  - Model-based



PURDUE UNIVERSITY
Department of Computer Science

Visual methods

- Box plot (1D)
- Scatter plot (2D)
- Limitations
  - Time consuming
  - Subjective



PURDUE UNIVERSITY
Department of Computer Science

Distance-based approaches

- Three major types of methods
  - Nearest-neighbor
  - Density-based
  - Clustering approach

PURDUE UNIVERSITY
Department of Computer Science

Nearest-neighbor

- Compute distance between every pair of points
- How to define outliers?
  - Points for which there are fewer than *p* neighboring points within distance *d*
  - Top *p* points whose distance to k^{th} nearest neighbor is greatest
  - Top *p* points whose average distance to their *k* nearest neighbors is greatest

PURDUE UNIVERSITY
Department of Computer Science

Example

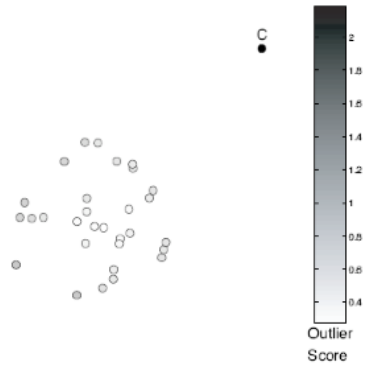


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

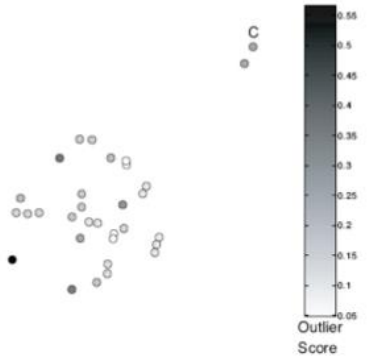


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

PURDUE UNIVERSITY
Department of Computer Science

Example

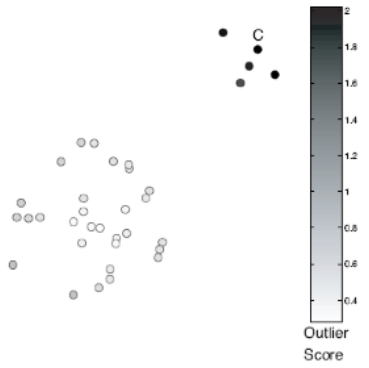


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

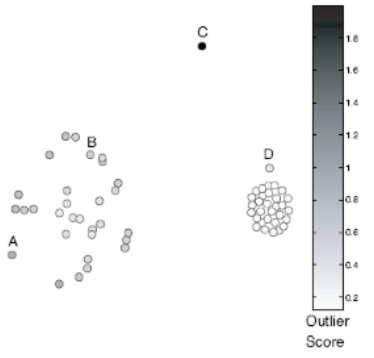


Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

PURDUE UNIVERSITY
Department of Computer Science

Density-based

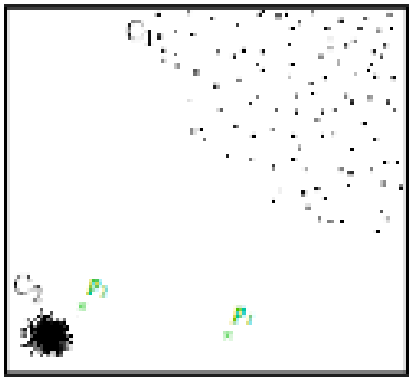
- For each point, compute the density of its local neighborhood of *k* neighbors

$$density(x,k) = \frac{|N(x,k)|}{\sum_{y \in N(x,k)} distance(x,y)}$$

- Local outlier factor (LOF) is the ratio of a point's density to the average density of its nearest neighbors

$$LOF(x) = \frac{\frac{1}{k}\sum_{y \in N(x,k)} density(y,k)}{density(x,k)}$$

- Outliers are points with largest LOF value



In the NN approach, p_2 is not discovered as an outlier, while the LOF approach considers both p_1 and p_2 to be outliers

PURDUE UNIVERSITY
Department of Computer Science

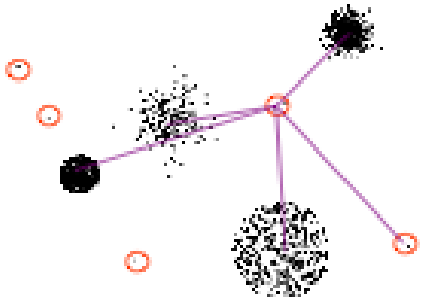
High dimensions

- In high-dimensional space, data is sparse and notion of proximity becomes meaningless
  - Every point is almost equally good outlier from the perspective of proximity-based definitions
- Lower-dimensional projections can be used for outlier detection
  - A point is an outlier if in some lower dimensional projection, it is present in a local region of abnormally low density

PURDUE UNIVERSITY
Department of Computer Science

Clustering-based

- 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



PURDUE UNIVERSITY
Department of Computer Science

Example

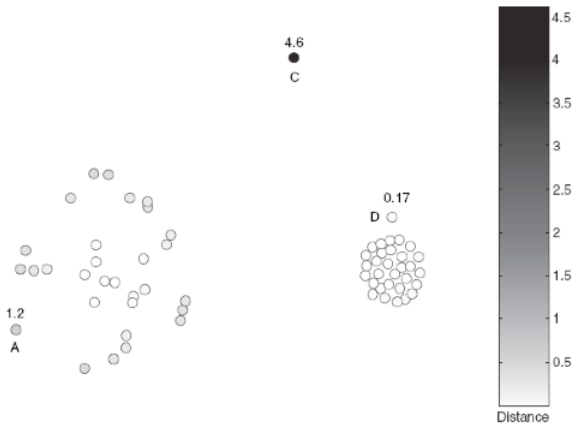


Figure 10.9. Distance of points from closest centroid.

This doesn't take into account the average distance of points to their cluster centroid (which can vary by cluster density)... so use relative distance (ratio of distance to median distance)

PURDUE UNIVERSITY
Department of Computer Science

Example

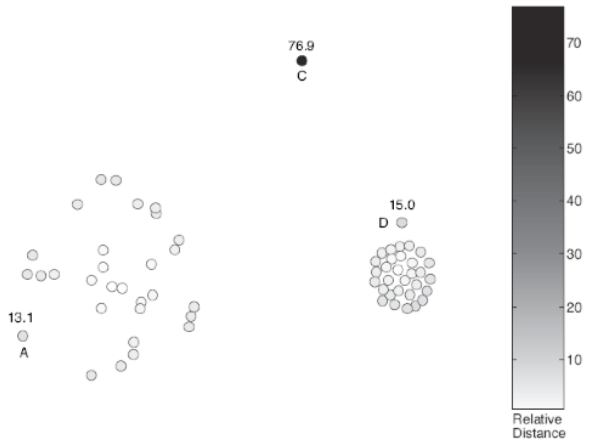


Figure 10.10. Relative distance of points from closest centroid.

PURDUE UNIVERSITY
Department of Computer Science

Supervised Anomaly Detection

- *Anomalies* not known in advance
  - Otherwise they wouldn't be anomalies
- But what if we assume we know normal?
  - Training data is from non-anomalies
- Train classifier to recognize normal
- One-Class Classification *(Moya & Hush '96)*

PURDUE UNIVERSITY
Department of Computer Science

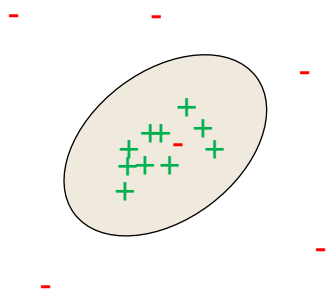
One-class Classification

- Problem: "Easy" to classify training data
  - Given instance, classify as normal
  - Works for all the training data
- Ideas:
  - "Fake" training data
  - Unpruned classifier
    - Narrowly tailor to recognize positive instances

PURDUE UNIVERSITY
Department of Computer Science

Artificial Training Data

- Start with a large volume of "normal" data
- Randomly generate a comparatively small volume of presumed anomalies
  - From uniform distribution across space
- Classifier accepts some false positives

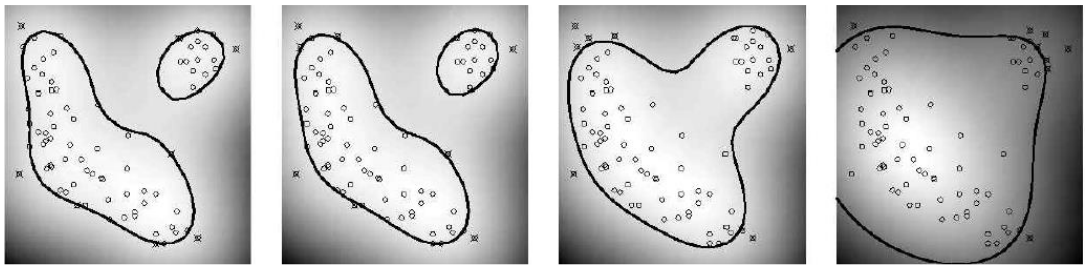


PURDUE UNIVERSITY
Department of Computer Science

One-Class Support Vector Machine

Schölkopf, Platt, Shawe-Taylor, Smola and Williamson 2001

- Idea: "Max-Margin" between points and "everything else"
  - Accomplished through clever choice of kernel functions
  - Max margin between points and origin gives boundary separating points from "everything else"



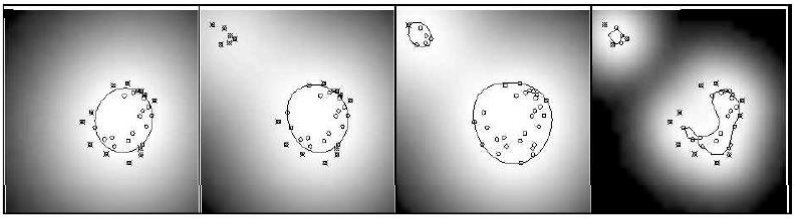
26

PURDUE UNIVERSITY
Department of Computer Science

One-Class Support Vector Machine

Schölkopf, Platt, Shawe-Taylor, Smola and Williamson 2001

- Parameters
  - ν bounds allowed outliers
  - γ governs classifier complexity (kernel function space)



ν, width c	0.5, 0.5	0.5, 0.5	0.1, 0.5	0.5, 0.1
frac. SVs/OLs	0.54, 0.43	0.59, 0.47	0.24, 0.03	0.65, 0.38
margin $\rho/\|w\|$	0.84	0.70	0.62	0.48

27

PURDUE UNIVERSITY | Department of Computer Science

Anomaly Detection: Statistical

Some materials from *Introduction to Data Mining* by Tan, Steinbach and Kumar

PURDUE UNIVERSITY
Department of Computer Science

Statistical approaches

- Use a parametric model to describe the data (e.g., Normal distribution)
- Apply a statistical test that evaluates how likely a point is under the data distribution
- Need to specify a confidence limit (e.g., 3σs away from mean is an outlier)

PURDUE UNIVERSITY
Department of Computer Science

Multivariate data

1. Statistical approach
   - Model data with a multivariate Gaussian
   - Calculate the likelihood of a point with respect to the estimated distribution, flag points with low likelihood as anomalous
2. Clustering approach
   - Use Mahalanobis distance to take into account the covariance of the attributes
   - Calculate distance of each point to the centroid, flag points with largest distance

$$d_{MH}(i,j) = \sqrt{\big(x(i) - x(j)\big)^T \Sigma^{-1} \big(x(i) - x(j)\big)}$$

PURDUE UNIVERSITY
Department of Computer Science

Example

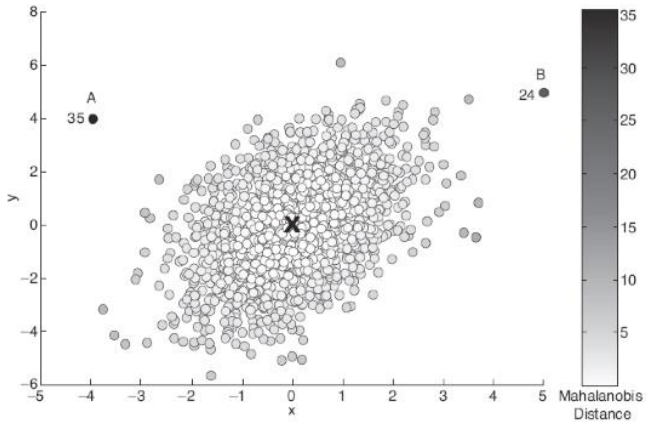
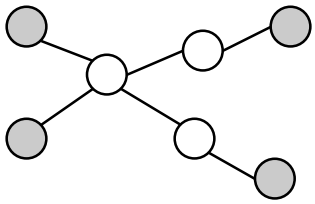


Figure 10.3. Mahalanobis distance of points from the center of a two-dimensional set of 2002 points.

PURDUE UNIVERSITY
Department of Computer Science

Example: Network traffic *(Lakhina et. al '04)*



Backbone network

Goal: Find **source-destination** pairs with high traffic (e.g., by rate, volume)

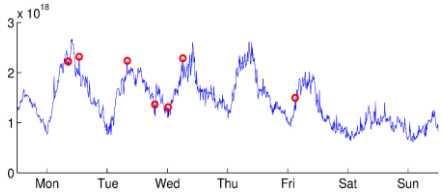
$$Y = \begin{bmatrix} \ldots. \\ 100 \;\; 30 \;\; 42 \;\; 212 \;\; 1729 \;\; 13 \\ \ldots. \end{bmatrix}$$

Source: Sutton, CS294 UCBerkeley

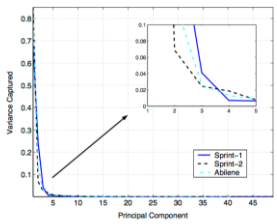
PURDUE UNIVERSITY
Department of Computer Science

Example: Network traffic

Abilene backbone network traffic volume over 41 links collected over 4 weeks



Perform PCA on 41-dim data
Select top 5 components to form "normal" subspace **P**



Source: Sutton, CS294 UCBerkeley

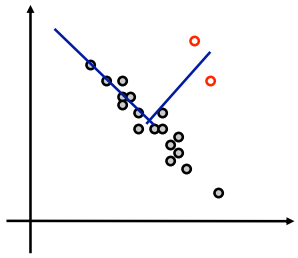
PURDUE UNIVERSITY
Department of Computer Science

Example: Network traffic

Data matrix

$Y = \begin{bmatrix} \ldots. \\ 100 \;\; 30 \;\; 42 \;\; 212 \;\; 1729 \;\; 13 \\ \ldots. \end{bmatrix}$

Perform PCA on matrix **Y**



Low-dimensional data

$Yv = \begin{bmatrix} \ldots. \\ y_t^T v_1 \;\; y_t^T v_2 \\ \ldots. \end{bmatrix}$

Eigenvectors

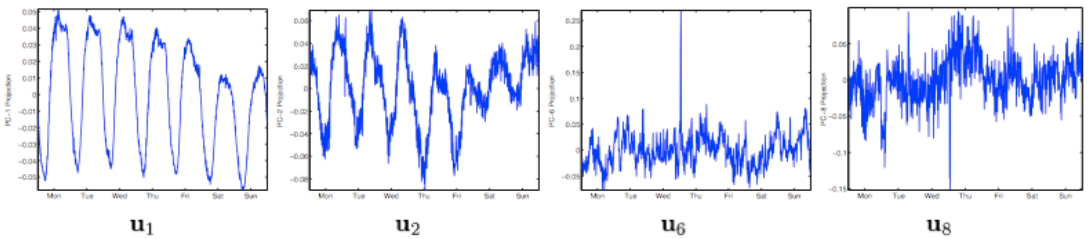
$\begin{bmatrix} v_1 \;\; v_2 \;\; \ldots \end{bmatrix}$

Source: Sutton, CS294 UCBerkeley

PURDUE UNIVERSITY
Department of Computer Science

Example: Network traffic

- Projections onto principal components: $u_i = \frac{\mathbf{Y}v_i}{\|\mathbf{Y}v_i\|}$
  - Look for first projection that contains a 3σ from mean to identify beginning of "anomalous" subspace



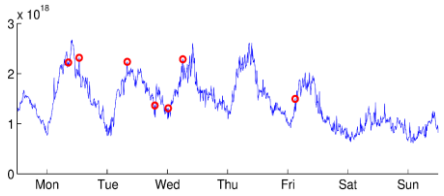
u_1 u_2 u_6 u_8

(a) Normal Behavior (b) Anomalous Behavior

PURDUE UNIVERSITY
Department of Computer Science

Example: Network traffic

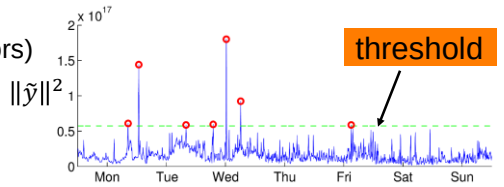
Abilene backbone network traffic volume over 41 links collected over 4 weeks



Project to residual subspace (everything but top 5 eigenvectors)

$$\tilde{\mathbf{y}} = \left(\mathbf{I} - \mathbf{P}\mathbf{P}^{\mathrm{T}}\right)\mathbf{y}$$

Flag anomalies using norm of projected vector



Source: Sutton, CS294 UCBerkeley

PURDUE UNIVERSITY
Department of Computer Science

Anomaly detection (generally)

- Knowledge representation?
  - **Categorization of data points as normal/anomalies**
- Pattern space?
  - **Set of data points** (each point is given a score)
- Score function?
  - **Distance from mean/centroid**
- Search?
  - **Exhaustive search**
    consider all points, return those with distance > threshold

38