MACHINE LEARNING & PUBLIC POLICY

LECTURE 3: ANOMALY DETECTION

Professor Edward McFowland III

Information Systems and Decision Sciences
Carlson School of Managment

University of Minnesota

PREDICTION FOR POLICY

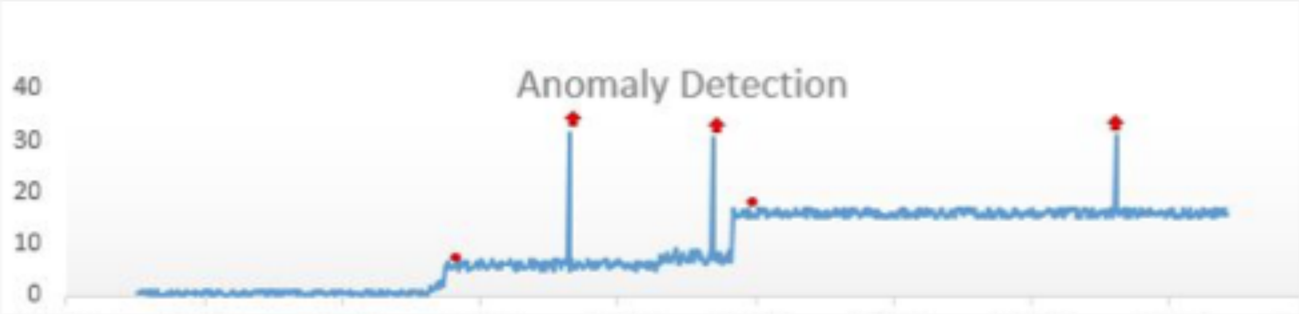
Observation 1: Sometimes correlation is valuable on its own

ANOMALY DETECTION

Observation 11: Sometimes goal is one of detection or discovery

ANOMALY DETECTION PARADIGM

- Identifying when a "system" deviates away from its expected behavior.



ANOMALY DETECTION

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data (anomalies are generated by a "different mechanism")
  - Interesting issue: anomaly detection vs. outlier detection
- Variations of anomaly/outlier detection problems
  - Given dataset D, find all data points $x \in D$ with anomaly scores $f(x)$ greater than some threshold t
  - Given dataset D, find all data points $x \in D$ having the top-n largest anomaly scores $f(x)$
  - Given dataset D, containing mostly normal (but unlabeled) data points and test point x, compute anomaly score $f(x)$ with respect to D

ANOMALY DETECTION

Main goal: focus the user's attention on a potentially relevant subset of the data.

1. Automatically **detect** relevant individual records, or groups of records.
2. **Characterize** and **explain** patterns: pattern type, affected subset, models of normal/abnormal data.
3. Present the pattern to the user.

Some common detection tasks

- Detecting **anomalous** records or groups
- Discovering **novelties** (e.g. new drugs)
- Detecting **clusters** in space or time
- Removing **noise** or **errors** in data
- Detecting **specific patterns** (e.g. fraud)
- Detecting emerging **events** which may require rapid responses.

EXAMPLES

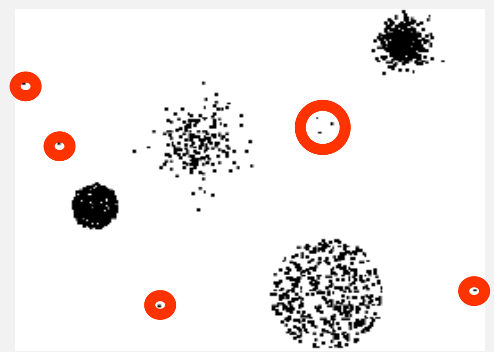
- Given a massive database of financial data, which transactions are suspicious and likely to be **fraudulent**?
- Given the huge number of container shipments arriving at our country's ports every day, which should be opened by customs (to prevent smuggling, terrorism, etc.)?
- Given a log of all the traffic on our computer network, which sessions represent (attempted) **intrusions**?

ANOMALY DETECTION

- Challenges?
  - How many outliers are there in the data?
  - Method is unsupervised
    - Validation can be quite challenging
  - Finding needle in a haystack
- Working assumption:
  - There are considerably more "normal" observations than "abnormal" observations (outliers/anomalies) in the data

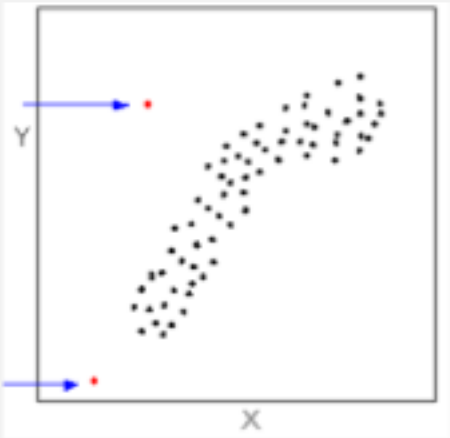
ANOMALY DETECTION

- General Steps
  - Build a **profile** of the "normal" behavior
    - I.e., patterns or summary statistics for the overall population
  - Use the "normal" profile to detect **anomalies**
    - I.e., observations whose characteristics differ significantly from the normal profile
- Detection schemes
  - Graphical
  - Distance/proximity - based
  - Statistical/model-based

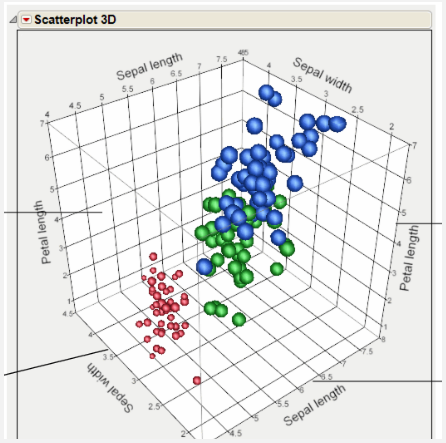


GRAPHICAL APPROACHES: EXAMPLES

SCATTER PLOT



3D SCATTERPLOT



DISTANCE BASED ANOMALY DETECTION

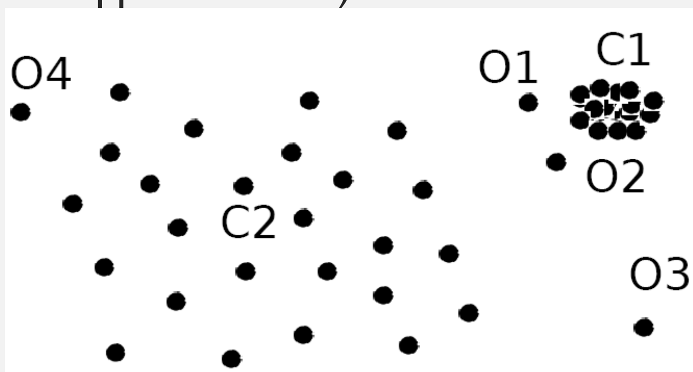
Observation II: Sometimes goal is one of detection or discovery

NEAREST-NEIGHBOR-BASED APPROACH

- Simple idea:
  - Compute the distance between every pair of data points and use the information about k nearest neighbors of each point
- There are various ways to define outliers:
  - Data points for which there are fewer than k neighboring points within distance d
  - Top n data points whose distance to the k-th nearest neighbor is greatest
  - Top n data points whose average distance to the k nearest neighbors is greatest

DENSITY-BASED APPROACH

- Finds **local outliers**, i.e., by comparing data points to their local neighborhoods, instead of looking at the global data distribution
- **Intuition**: The density around an outlier object is significantly different from the density around its neighbors
- **Method**: Use the relative density of an object against its neighbors as the indicator of the degree of the object being outliers
- **Example**: O1 and O2 are local outliers (to C1), O3 is a global outlier, but O4 is not an outlier. Nearest-neighbor-based approaches would not identify O1 and O2 as outlier (as opposed to O4).

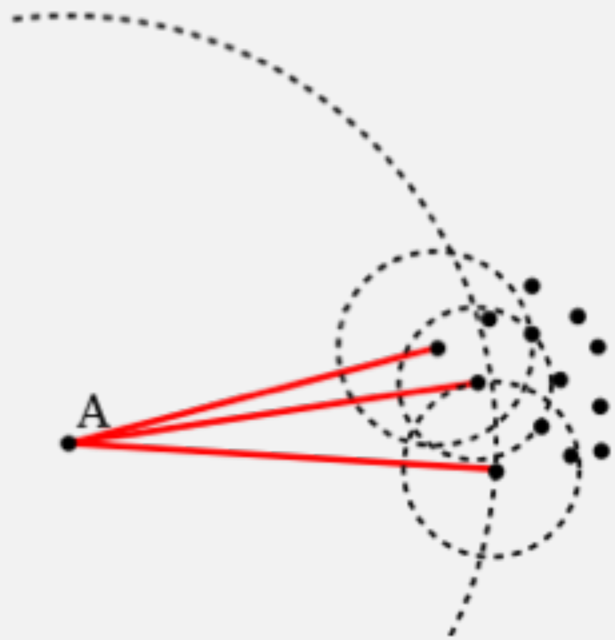


DENSITY-BASED APPROACH: LOCAL OUTLIER FACTOR (LOF)

- Basic idea:
  - For each object (data point), compute the density of its local neighborhood (defined by the *k* nearest neighbors)
  - Compute local outlier factor (LOF) of a given object as the ratio between its local density and the local densities of its nearest neighbors
  - Outliers are objects with largest LOF value
- A number of further variations and refinements have been proposed

LOF APPROACH: EXAMPLE

- Object A has much lower local density than its nearest neighbors



Source: wikipedia.org

LOF APPROACH: DETAILS

- **k-distance** of object *A*, ***$dist_k(A)$***
  - Distance between *A* and its k^{th} nearest neighbor
- **k-distance neighborhood** of *A*, ***$N_k(A)$***
  - $N_k(A) = \{B \mid B \in D, dist(A,B) \leq dist_k(A)\}$
  - Esentially, $N_k(A)$ is the set of k nearest neighbors of A
  - However, technically size of $N_k(A)$ could be bigger than *k* since multiple objects may have identical distance to *A*
- **Reachability distance** of *A* from *B*: ***$reachdist_k(A,B)$***
  - $reachdist_k(A,B) = \max \{ dist_k(B), dist(A,B) \}$
  - I.e., objects *A* that belong to the *k* nearest neighbors of *B* have the same $reachdist_k(A,B)$

LOF APPROACH: DETAILS (2)

- **Local reachability density** of A, **lrd(*A*)**
  - $\text{lrd}(A) = 1 / \left(\sum_{B \in N_k(A)} \textit{reachdist}_k(A,B) / |N_k(A)| \right)$
  - I.e., captures how *A* can be reached from its neighbors

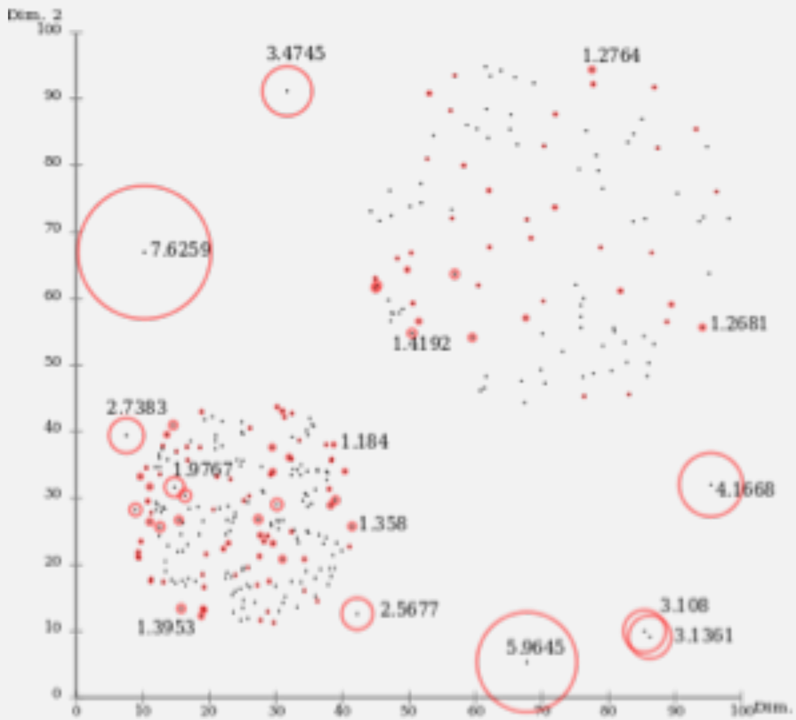
- **Local outlier factor** of A, ***LOFk(A)***

$$\text{LOF}_k(A) := \frac{\sum_{B \in N_k(A)} \frac{\text{lrd}(B)}{\text{lrd}(A)}}{|N_k(A)|} = \frac{\sum_{B \in N_k(A)} \text{lrd}(B)}{|N_k(A)|} / \text{lrd}(A)$$

  - I.e., average local reachability density of A's neighbors divided by the A's own local reachability density

LOF APPROACH: EXAMPLE (2)

- LOF(*x*) = 1: data point *x* is comparable to its neighbors (not an outlier)
- LOF(*x*) < 1 indicates a denser region
- LOF(*x*) significantly larger than 1 indicate outliers



Source: wikipedia.org

MODEL BASED ANOMALY DETECTION

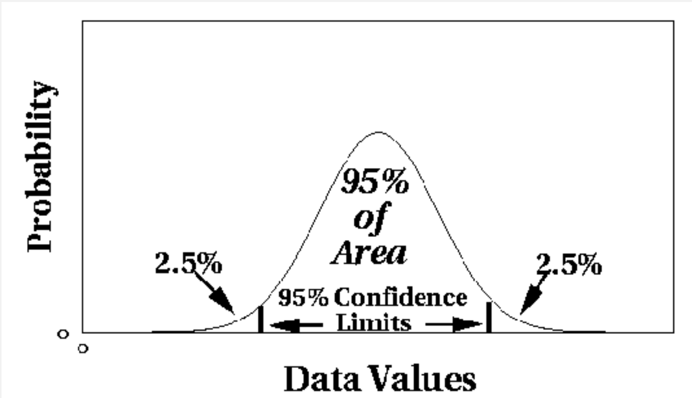
Observation 11: Sometimes goal is one of detection or discovery

STATISTICAL APPROACHES

- Statistical methods (also known as model-based methods) assume that the regular data follow some statistical model (a stochastic model)
  - The data not following the model are outliers
  - Lots of different models are available
- Effectiveness of statistical methods highly depends on whether the assumption of statistical model holds in the real data
  - Many statistical techniques have been developed
  - E.g., parametric vs. non-parametric

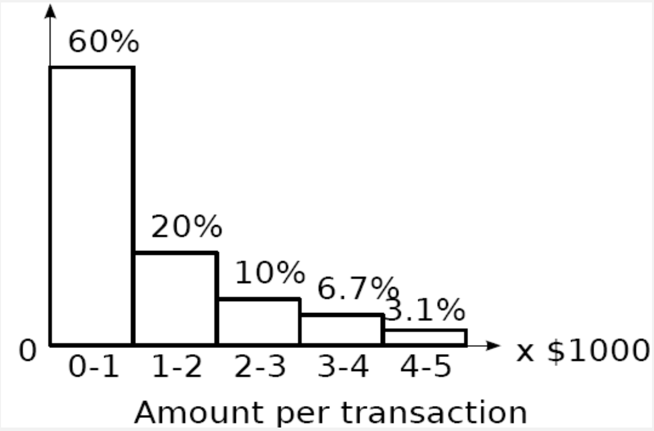
STATISTICAL APPROACHES: GENERAL IDEA

- Assume a parametric model describing the distribution of regular data (e.g., normal distribution)
- Apply some statistical test/procedure on how likely is that a given data point was generated by the assumed distribution



NON-PARAMETRIC METHODS FOR ANOMALY DETECTION

- **Non-parametric**: The model of regular data is learned from the input data without any *a priori* structure
- Fewer assumptions about the data – applicable in more scenarios
- Example:
  - Histogram-based approach



References

- A coherent text on anomalous pattern detection has yet to be written, but many methods have been proposed and are becoming common:

  - WSARE: W.-K. Wong et al., "Rule-based anomaly pattern detection for detecting disease outbreaks," *Proc. 18th Natl. Conf. on Artificial Intelligence*, 2002.
  - APD ("Anomaly Pattern Detection"). K. Das, J. Schneider, and D.B. Neill, *Proc. KDD 2008*.
  - D.B. Neill and W.-K. Wong, "A Tutorial on Event Detection," presented at *KDD 2009* conference.
  - Edward McFowland III, Skyler Speakman, and Daniel B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14: 1533-1561, 2013.

- Software for spatial cluster detection and for WSARE is available on the Auton Laboratory web page, http://www.autonlab.org.